資料探勘 mini project 梁錦順 40947055S

資料集說明:
這個資料集包含了關於患者的多個特徵，這些特徵可以用來預測患者是否患有糖尿病。

特徵說明:

p_id	患者 id
no_times_pregnant	懷孕次數
glucose_concentration	口服葡萄糖耐量測試中 2 小時的血漿葡萄糖濃度
blood_pressure	舒張壓（毫米汞柱）
skin_fold_thickness	肱三頭肌皮膚褶厚度（毫米）
serum_insulin	2 小時血清胰島素濃度（mu U/ml）
bmi	身體質量指數（體重（公斤）/身高（米）^2）
diabetes pedigree	糖尿病家族史函數
age	年齡（歲）
diabetes	類別變數（0 或 1）

前處理:

- p_id只是為了標註患者，每個id都是獨特的 對模型學習沒有幫助所以不會包含
- 由於test.csv沒有diabetes欄位 為了能計算accuracy我先採用random sampling來分開資料成training跟test 比例為（80% training : 20% test）
- 有嘗試將資料normalize（min max normalization）但是實驗結果很悲觀（accuracy 很低 0.454）所後來沒有採用

分類器:

- knn
- random forest classifier

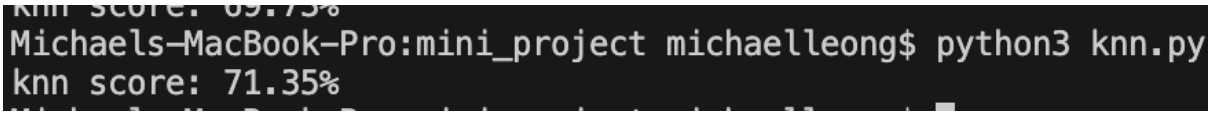
knn (knn.py):
計算score如下:

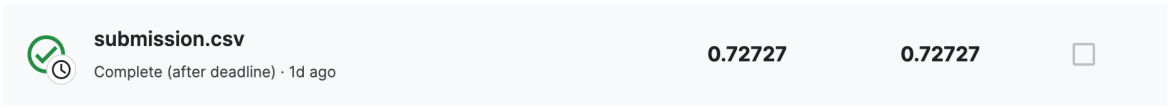
```
print("knn score: {:.2f}%".format(knn.score(x_test, y_test)*100))
```

我先設k為2當初也是採用了normalization 結果如下圖

```
4   181           1
Michaels-MacBook-Pro:mini_project michaelleong$ python3 main.py
knn score: 69.19%
```

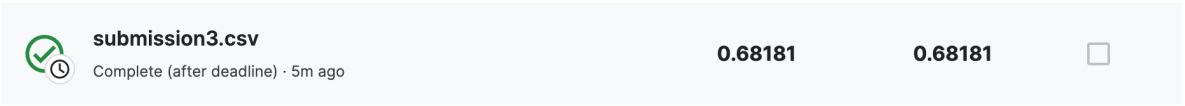
我後來不採用normalization 結果如下圖





能看到accuracy提高的非常多

我接著嘗試把k設為3 結果如下圖

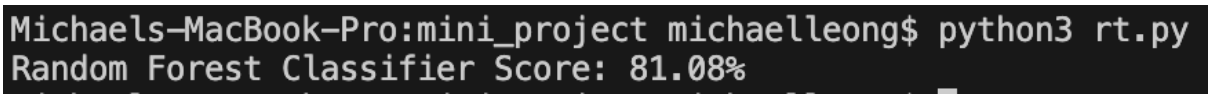


accuracy降了;結論k設為2比較好

random forest classifier (rt.py):
計算score如下:

```
print("Random Forest Classifier Score: {:.2f}%".format(rt.score(x_test, y_test)*100))
```

我嘗試用別的分類器我設定max_depth=5,min_samples_split=3結果如下圖



結果比使用knn還要高

最終成績(選kaggle最高分)(submission2.csv):