專題實作目的

探討特徵篩選、分類器集成策略 (random forest) 結果

採用 data mining 模組

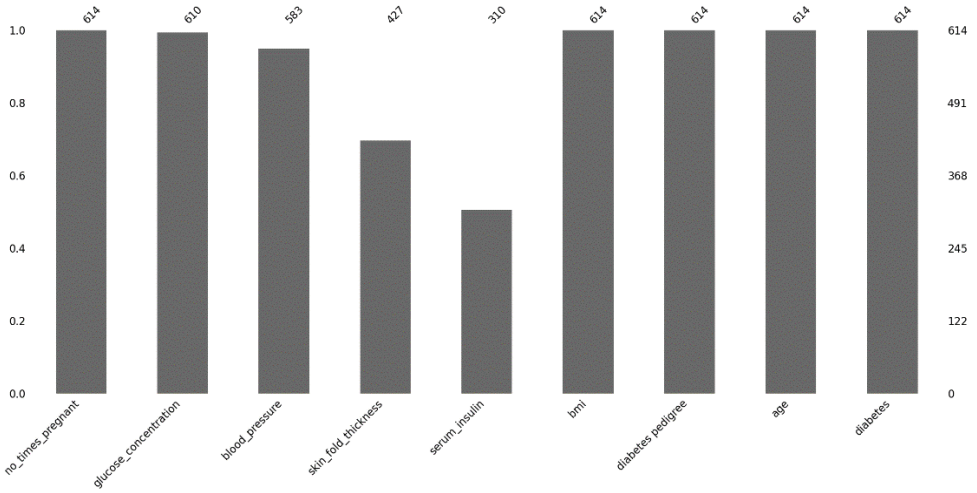
Python scikit-learn

程式/環境設定

Python3 和 Jupyter Notebook 開發環境

改變控制參數說明

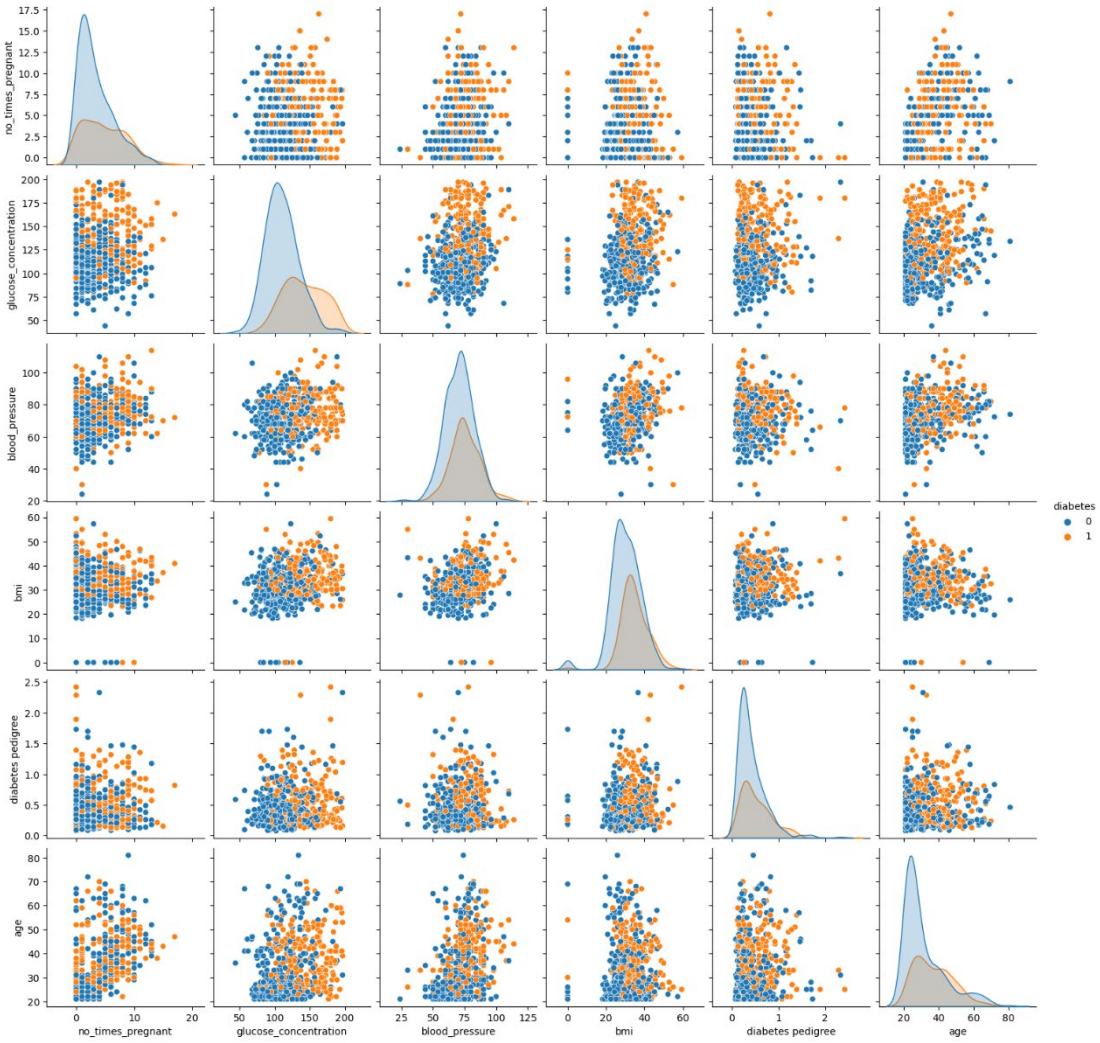
1. 前處理
   去除 p_id 後，印出長條圖觀察數據缺失值，發現 skin_fold_thickness 和 serum_insulin 缺失值過多，決定 drop 掉



   但觀察到 glucose_concentration 和 blood_pressureuum 也有缺失值，

考慮到這兩中資料應是連續型的，所以採取線性插值方式補齊缺失值

原始資料分布



調整後數據分布

可以看見修正後，資料更為集中

2. 建立模型

一開始使用預設的參數建模，得到 AUC = 0.70，及各項特徵重要程度

(按懷孕次數、葡萄糖濃度、血壓、BMI、糖尿病指數、年齡)

無限制

```
0.7058139534883721
特徵重要程度: [0.08501015 0.31704809 0.10294974 0.18088225 0.15632677 0.15778301]
```

再套用常見 RF 參數(n_estimators = 100, max_depth = 5)，得到 AUC =0.65

```
0.6539244186046512
特徵重要程度: [0.07321357 0.43114759 0.05858259 0.17604448 0.11573197 0.1452798 ]
```

3. 修正

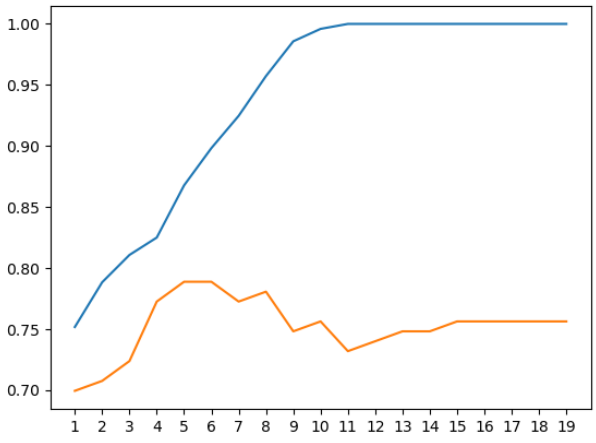
注意到懷孕次數的特徵重要度都在 0.1 之下，思考或許為干擾因素，因此將其 drop 掉，觀察變化，得到 AUC = 0.76，且反覆測試幾次都在 0.7 之上。

(特徵依排序是: 葡萄糖濃度、血壓、BMI、糖尿病指數、年齡)

```
去掉懷孕次數:
0.7675872093023256
特徵重要程度: [0.43815229 0.06650255 0.1963523  0.13362244 0.16537042]
```

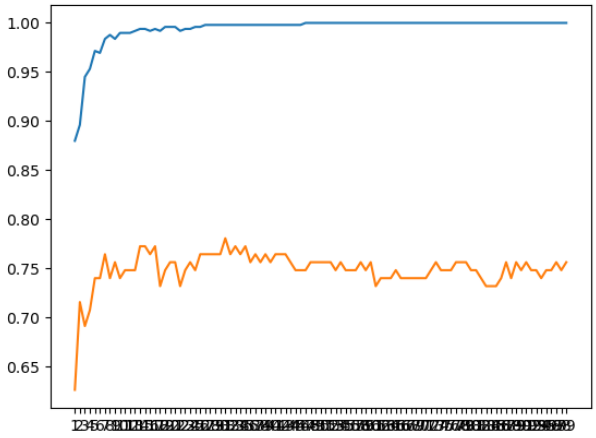
1. 再畫出

Accuracy vs depth(depth = 1 ~ 20)




```
max acc: 0.7886178861788617
5
```

Accuracy vs estimator(estimator = 1~100)




```
max acc: 0.7804878048780488
31
```

4. 最終模型

用 Max_depth = 5, n_estimator = 31 避免 overfitting，得到 AUC = 0.79

```
最終模型:
0.7934593023255814
特徵重要程度: [0.36518237 0.07320272 0.22046237 0.1408414  0.20031114]
```

評估方式

2. AUC score
3. Accuracy vs estimator
4. Accuracy vs depth

結果與討論

最終結果得到 AUC = 0.79，算是個差強人意的結果，看完同學們 demo 的準確度都有達 0.9 以上的分數，難免有點灰心。模型的修正已盡量避免 overfitting 的狀況，可能是因資料量過少(train + test 共 768 筆)，才讓 random forest 分得沒那麼細；另一方面可能是前處理方式過於粗糙，可以從上面處理完的分布圖中看出，離群值還是不少，而且直接刪除 3 種 feature 也可能刪除很多有用的資訊。覺得可以改進插值(目前用線性插值)的方法，找到更貼近現實狀況的插值方式，這樣那些缺失值可以補足，就能多用 2 種 feature。血壓那項，或許可以依據低血壓、正常、高血壓，先做數值 normalize，讓資料更為集中、簡單化。