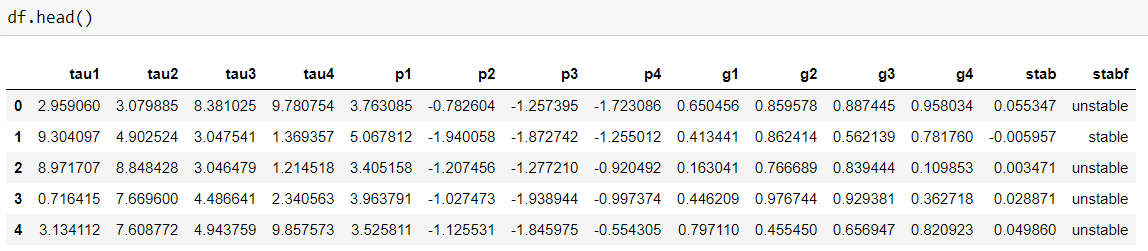
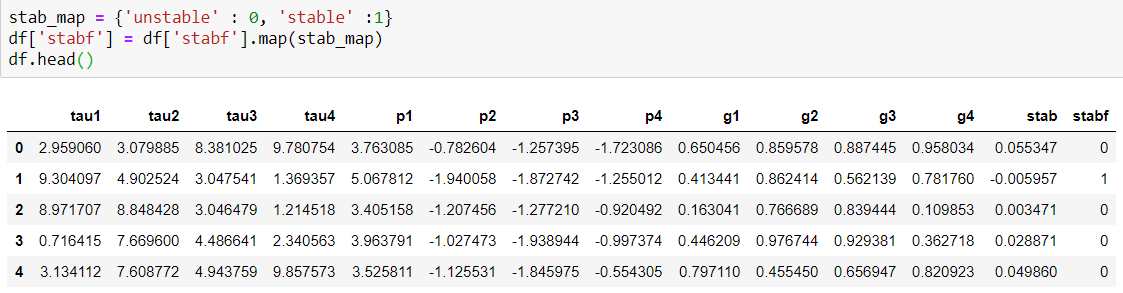
**資料分析與學習基石HW4**

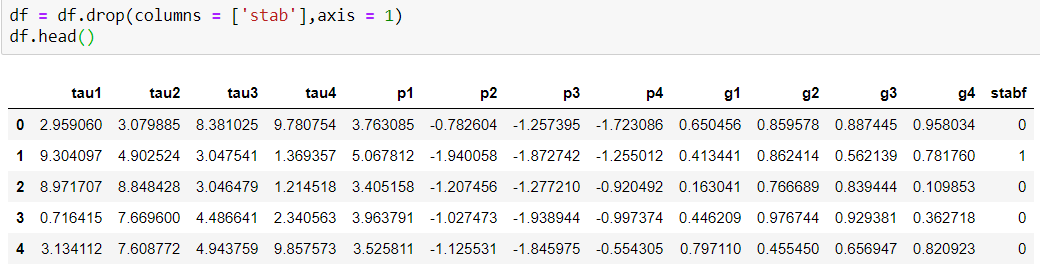
1. Choose a dataset: “Electrical Grid Stability Simulated”
2. **分析各個參數對電壓是否穩定的影響，並探討何種model比較準確**
3. Analyze the data



更改stabf欄位，unstable = 1, stable = 0



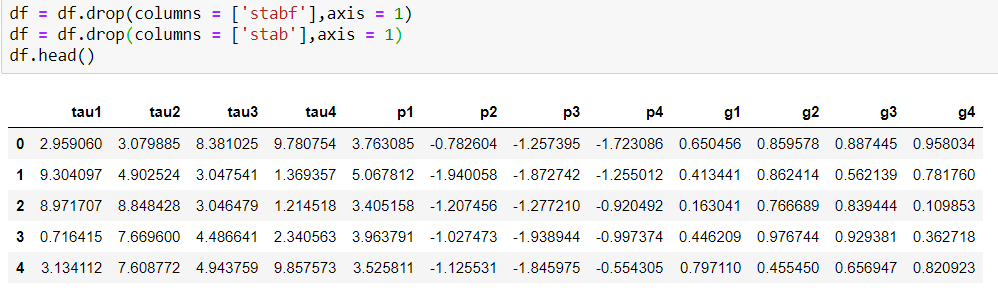
為了不干擾stabf所以除去stab欄位



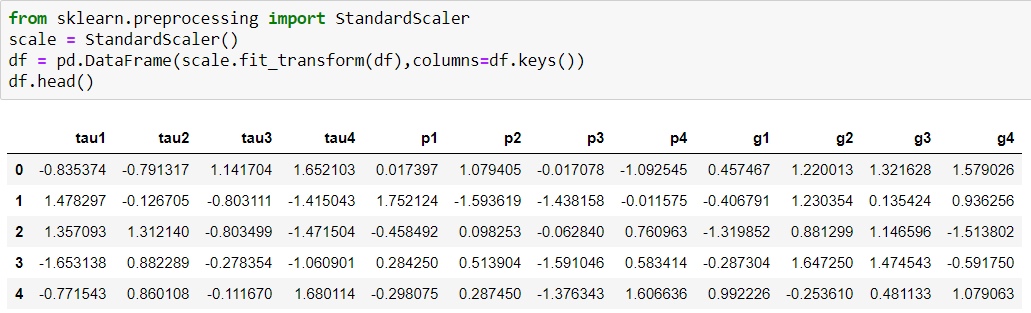
運用seaborn的熱圖觀察各參數對穩定度之相關係數



將要拿來預測的數據分離出來



接著標準化



完成數據前處理

1. **開始預測**

我選了4個模型LogisticRegression，SVM，DecisionTree，RandomForest來做比較。

而分割train，test我一開始是用隨機切割，但是後來發現隨機切割會造成每次預測的結果有不同，拿來對比會失真，所以用手動分割(7:3)以方便比較，但是實測結果是隨機分割之準確率較高，高2%左右。以下結論為手動分割準確率

**Original Result:**

SVM(0.9517) > RandomForest (0.879)> DecisionTree(0.836) > Logistic Regression(0.8036)

**Reasons:**

1. 由於Logistic Regression與SVM較相似所以拿來比較， Logistic Regression 與SVM不同的點是 LR不用像SVM要找最佳距離，而SVM對於異常數據比較敏感，因為training時只需要找支持向量，一旦被干擾，預測結果就會不如預期。
2. 而Logistic Regression在加入更多特徵時期準確率變低是在於，LR不擅長處理大的特徵空間，而SVM反之，所以在加入更多特徵時準確率提升。

結合(1)(2)，本例說明當特徵空間大但不是過大時時，SVM基本上會優於LR，前提是無太多異常數據干擾SVM，這個例子是電廠，不太會有異常數據，主要都是由公式推得，而預測結果也證明如此。

Random Forest:

隨機森林能一定程度減少決策樹的overfitting，而樣本隨機特徵隨機，再經過多樣本的訓練，最後預測結果較好，在處理多特徵時是不錯的選擇，但訓練過程不透明是缺點，準確率排名第2。

Decision Tree:

決策樹比較能夠處理不相關的數據但是容易忽略數據之間的相關性，且易發生overfitting，加上其結果會偏向那些具有更多數值的表徵。但是本例的參數(tau，p，g)一樣多，只有P的參數之間比較有相關性，tau跟g之間比較無關所以整體來說準確率不是特高但還算不錯。所以會看到在拿掉P時，準確率由0.836上升至0.852，與決策樹性質吻合。

**Approaches & Improvement:**

藉由觀察熱圖，發現P1、P2、P3、P4這4個參數對結果的相關係數最小，所以我先試了幾種組合(拿掉所有P，拿掉P2P3、拿掉P1P2P3、拿掉P2P3P4)

結果如下圖



大致上來說，拿掉P確實對提升準確率有些許幫助，但是不同的P排列組合的影響非常小，影響不到1%。

最後，在使用SVM模型及拿掉最低相關係數的前三名P1P2P3時準確率達最高。

而P的排列組合影響沒那麼應該是因為P是由簡單的加法公式得出，

P1 = abs(P2+P3+P4)，彼此之間呈高度負相關(因為P1是絕對值)，而其他參數之相關係數太相近所以我就沒動。