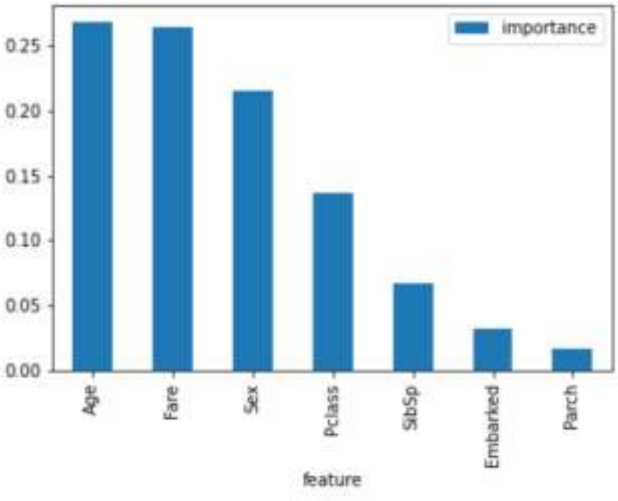
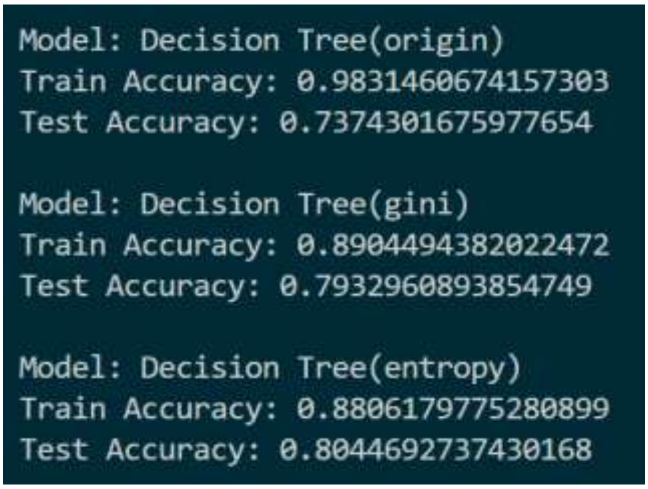
我把除了 embarked 外的資料都拿去 train，在其他都沒改的情況下 test accuracy 變成 0.7430167597765364，就已經符合標準了，但是 training accuracy 卻有 0.9859550561797753，代表 overfitting 了。我在網路上有找到一張圖表，他分析了 training data 中那些資料比較重要，我覺得除了原本的 Sex, Age, Fare 外，可以再加入 Pclass。



決策樹的改進:

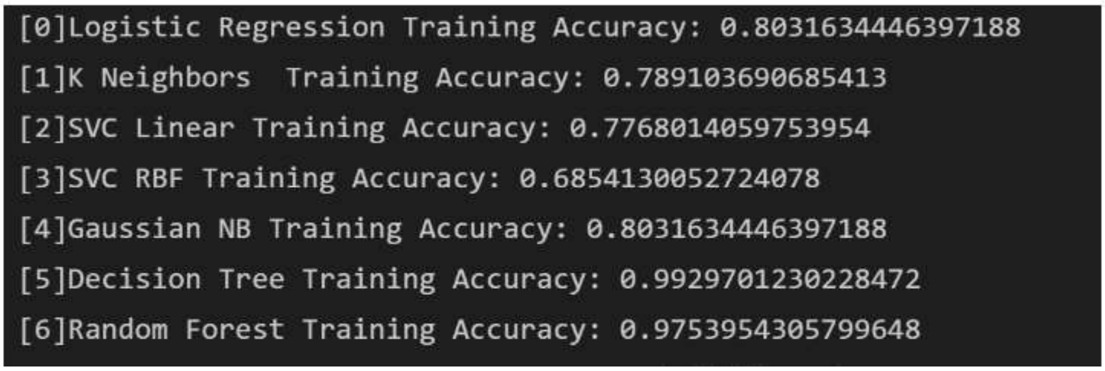
1. Criterion :用 gini 或 entropy 其實相差不大
2. Splitter: 我選擇用 random 比較可以防止 overfitting
3. Max_depth:我發現 10 大概是可以讓 test accuracy 比較高



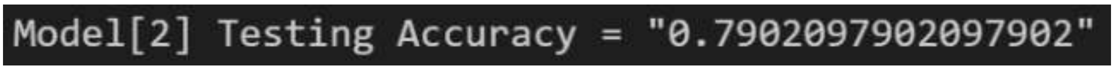
```
Model: Decision Tree(origin)
Train Accuracy: 0.9831460674157303
Test Accuracy: 0.7374301675977654

Model: Decision Tree(gini)
Train Accuracy: 0.8904494382022472
Test Accuracy: 0.7932960893854749

Model: Decision Tree(entropy)
Train Accuracy: 0.8806179775280899
Test Accuracy: 0.8044692737430168
```

模型的選擇:

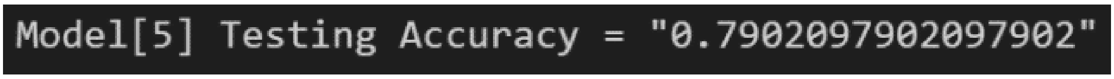
我上網有查到有人把一些模型一起用，看哪個模型表現較好



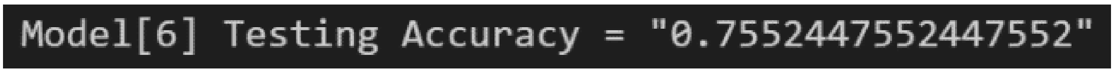
```
[0]Logistic Regression Training Accuracy: 0.8031634446397188
[1]K Neighbors  Training Accuracy: 0.789103690685413
[2]SVC Linear Training Accuracy: 0.7768014059753954
[3]SVC RBF Training Accuracy: 0.6854130052724078
[4]Gaussian NB Training Accuracy: 0.8031634446397188
[5]Decision Tree Training Accuracy: 0.9929701230228472
[6]Random Forest Training Accuracy: 0.9753954305799648
```



```
Model[2] Testing Accuracy = "0.7902097902097902"
```



```
Model[5] Testing Accuracy = "0.7902097902097902"
```



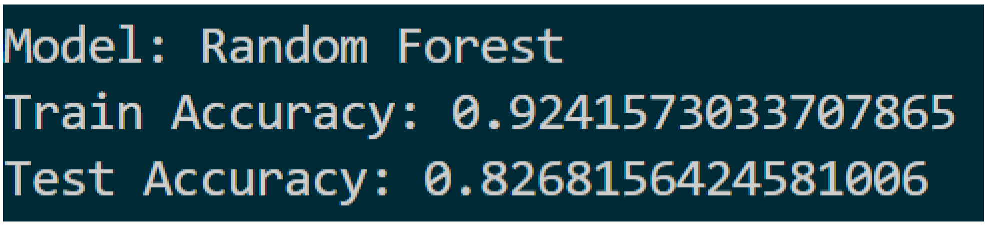
```
Model[6] Testing Accuracy = "0.7552447552447552"
```

所以我就選擇用 SVC 和 Random Forest

Randon Forest:用多個 Decision Tree 來分類，輸出的類別由個別樹的輸出眾數決定

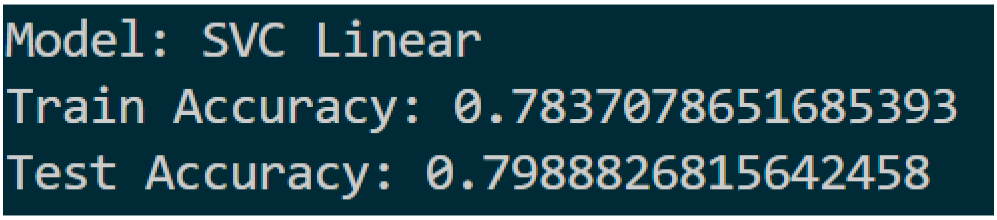
1. Boostrap 設為 True，表示隨機抽樣

n_estimators = 200, max_depth = 9,這是我試出來的結果會最好



```
Model: Random Forest
Train Accuracy: 0.9241573033707865
Test Accuracy: 0.8268156424581006
```

SVC:只有 linear 的模型可以有最高的 test accuracy



```
Model: SVC Linear
Train Accuracy: 0.7837078651685393
Test Accuracy: 0.7988826815642458
```