Data Mining - Project 2 Classification

電通甲 Q36061240 洗鈺淇

I. Dataset Introduction

Survival on the Titanic 鐵達尼號上的乘客資訊與是否倖存之關係

Kaggle 網址: https://www.kaggle.com/heptapod/titanic

Data size :891 條 x12 欄

欄位說明:

Colum	Type	Description
PassengerId	Numeric	乘客 ID
Survived	Srting, Discrete	是否倖存(S=倖存 C=死亡)
Pclass	Numeric, Discrete	乘客艙位等級(1-頭等艙 2-二等艙 3-三等艙)
Name	Srting	乘客姓名
Sex	Srting, Discrete	性別
Age	Numeric	年齡
SibSp	Numeric	手足與配偶個數
Parch	Numeric	父母與小孩個數
Ticket	Numeric	船票編號
Fare	Numeric	票價
Cabin	String	客艙座號(字母+數字組合)
Embarked	String, Discrete	登船港口

Table 1 Field Description

# PassengerId	# Survived	# Pclass	A Name	A Sex	# Age	# SibSp	# Parch	# Ticket	# Fare	A Cabin	A Embarked
1 … 891	0 … 1	1 … 3	891 unique values	male 65% female 35%	0.42 … 80	0 … 8	0 … 6	693 … 3.1m	0 … 512	147 unique values	S 72% C 19% Other (1) 9%
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S

Figure 1 Dataset Content

■ **初步統計分析:**

首先統計出在這份資料集中 891 人的各項特徵的分布圖，其中倖存人數約為 300 人，倖存率為 38.38%；依座艙等級來看人口分布的話，大多數人皆位於三等艙，而三個等級的座艙年齡分布的趨勢差不多，但發現頭等艙均年齡比較大，因為大多中年至老年人較富有，符合財富與年齡的趨勢走向；登船港口人數分布按 S、C、Q 遞減，我推測可能是 S 港口為比較大的港口才會比較多人從此港口上船，又或 S 港口坐落於人口密集區，交通發達，才可能符合這個統計。

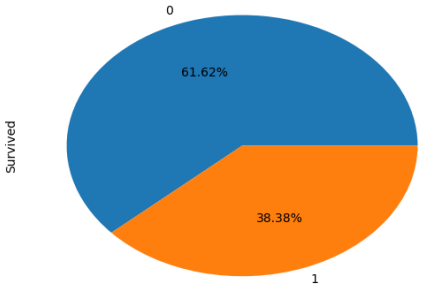


Figure 1 The fraction of survived

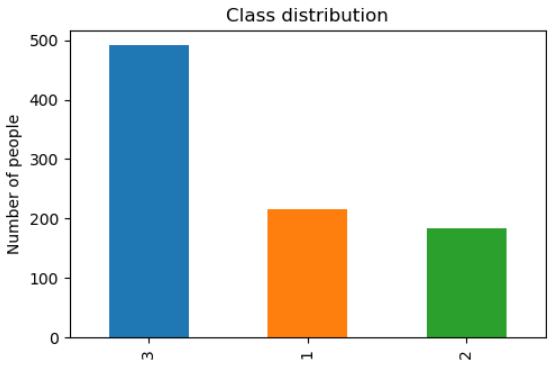


Figure 2 各艙等人口分布

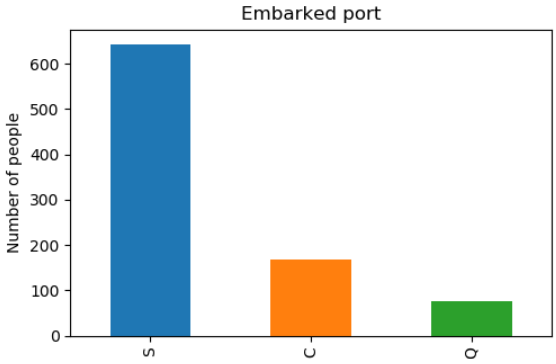


Figure 3 各登船港口人口分布

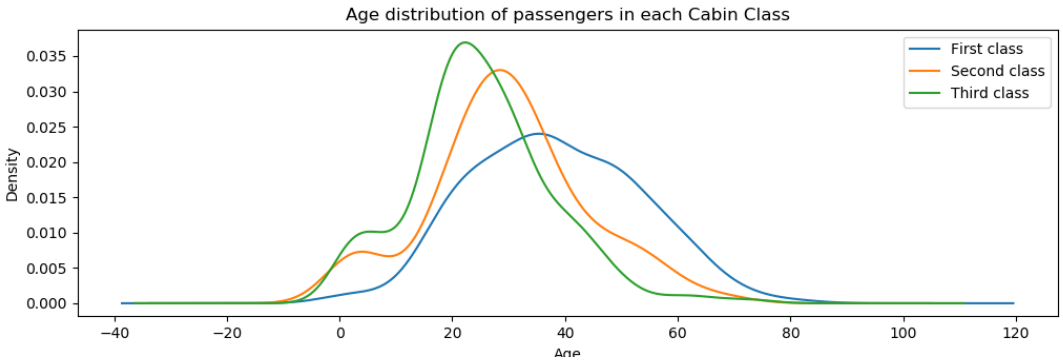


Figure 4 各艙等年齡分布

II. Data Analysis

接下來考慮各個特徵與倖存率之間的關係，依照所看過的鐵達尼號電影以及查詢相關新聞後，有發現到當時在船上的人會選擇優先保護女性以及小孩，讓他們先搭乘救生艇離開，也進一步推測是否會先救援比較有經濟地位的頭等艙乘客，這邊在用統計圖表來呈現。

(a) 性別與座艙等級

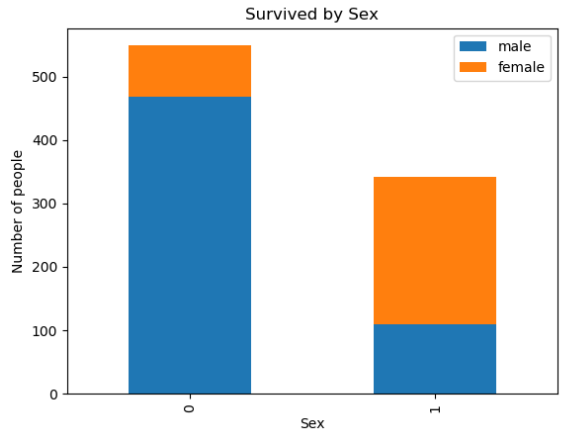


Figure 5 倖存人數的性別比率(0:死亡 1:倖存)

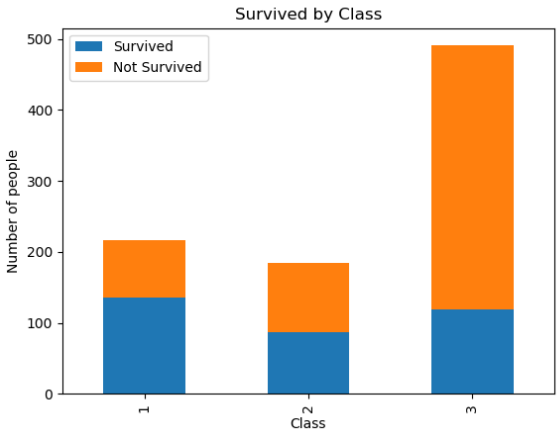
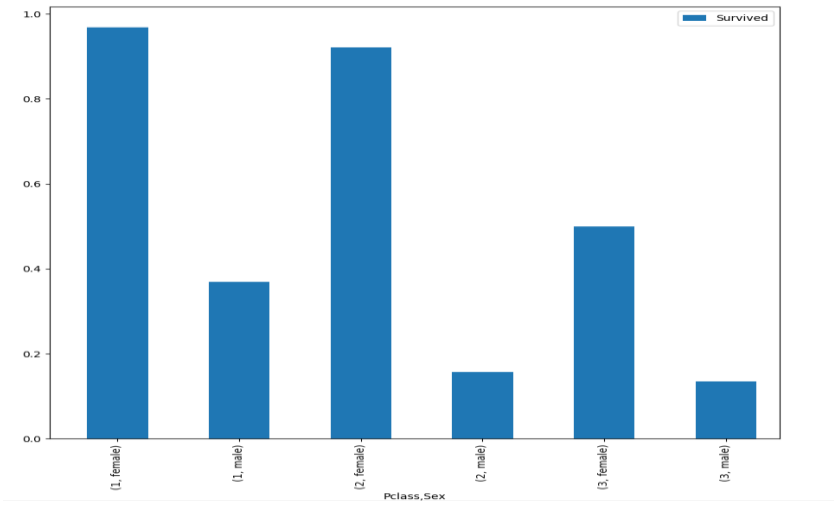


Figure 6 各艙等的倖存人數

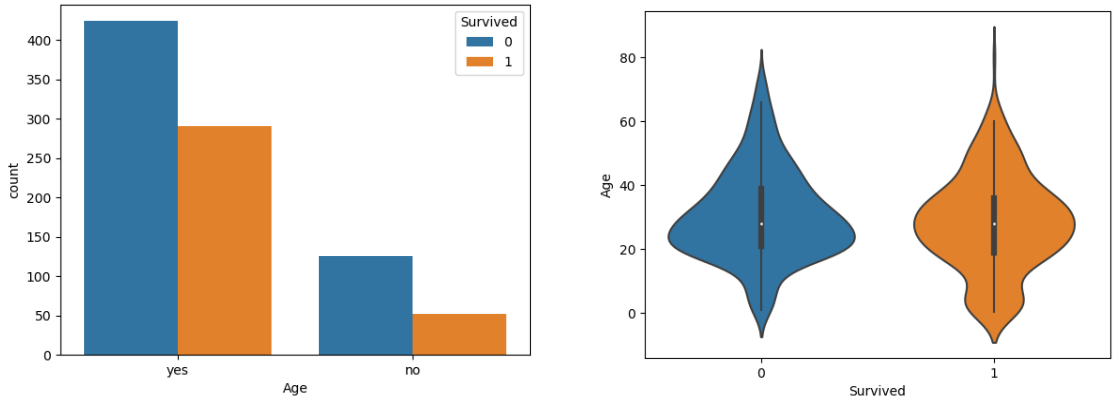
由圖表上可得知，女性獲救人數比男性多很多，而依照座艙等級來分的話，頭等艙的獲救人數最多，在依照各艙等的獲救人數的性別比例來看，頭等艙及二等艙的女性獲救比率相當高非常趨近於 1。



(b) 年齡

由於年齡是有 missing value 而且是連續型的變數，於是先看一下有年齡標記的乘客

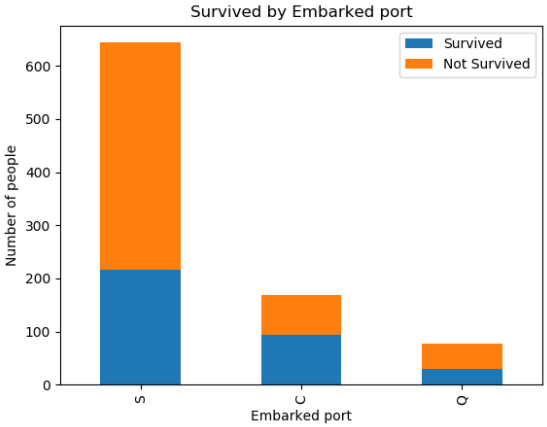
與沒有標記的獲救程度(1 為獲救)，由圖中顯示有標記的獲救率比較高(可能因為沒有標記的都已經不幸死亡)，而由年齡和獲救人數分布圖來看，比較明顯的部分是青少年及兒童獲救程度較高。



因此在前處理部分會將 Age 缺失的部分以平均值填滿和以年齡大小區分三組（Child:小於 12 歲/Youth:小於 20 歲/Adult:小於 60 歲/Old:大於 60 歲）來比較分類結果

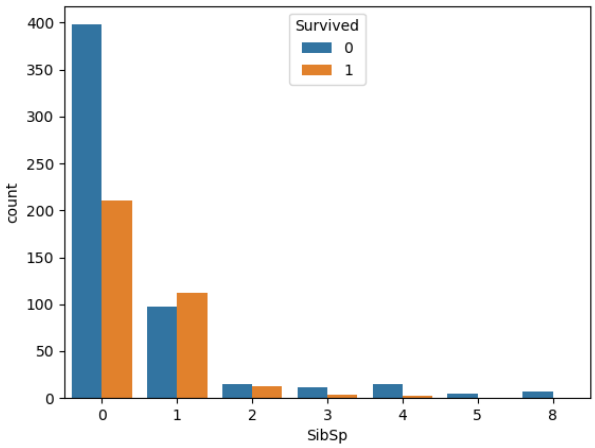
(c) 登船港口

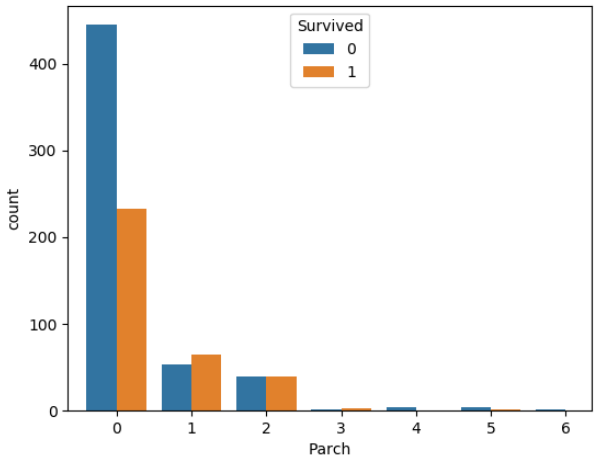
大多數的人都是從 S 港口登船，可知 S 港口是處於人口密集處，另外 C 港口的倖存率有達六成，推測是與該港口登船乘客的經濟地位有關。



(d) 手足配偶人數、父母兒女人數

由統計圖發現當手足配偶或父母兒女人數只有一人的時候，該乘客獲救率會超過一半，這可能反映當時會優先選擇家裡親人只有一位的先救援；而大部分死亡的乘客的親人都為 0 個，推測當時許多窮困移民大多都是無依無靠而會選擇讓家裡有親人的乘客先行搭乘救生艇。

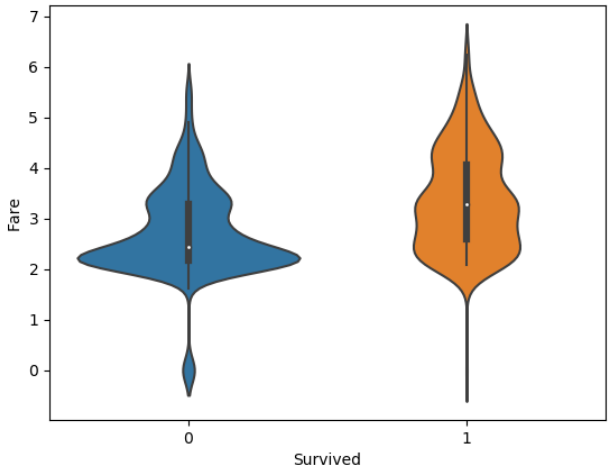




因此我們將這 SibSp 分三群依照 small: SibSp<1、middle:1<SibSp<3、large: SibSp>3，而 Parch 也是分三群 small: Parch<1、middle 1<Parch<3、large: Parch>3。

(e) 票價

票價的 range 很廣，於是取 log 來看，log(Fare)小於 2.5 時，死亡率是高於生存率的，而大於 2.5 的生存率是高於死亡率，因此我們把這個 feature 分成兩群 rich 為 log(Fare)>2.5，poor 為 log(Fare)<2.5。



III. Preprocess

- **Missing value**

發現了 Age、Cabin、Embark 有 missing value，其中又 Cabin(座位)的 missing value 最多，所以將其分成有紀錄的跟沒有的兩類，Age 的 missing Value 選擇填入均值，接著由於 Embark 僅有 2 個 missing value 則選擇去掉這兩條資料。

```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
```

Figure 7 處理前欄位資訊

```
Data columns (total 12 columns):
PassengerId    889 non-null int64
Survived       889 non-null int64
Pclass         889 non-null int64
Name           889 non-null object
Sex            889 non-null object
Age            889 non-null object
SibSp          889 non-null object
Parch          889 non-null object
Ticket         889 non-null object
Fare           889 non-null object
Cabin          889 non-null object
Embarked       889 non-null object
dtypes: int64(3), object(9)
```

Figure 8 處理後欄位資訊

	PassengerId	Survived	Pclass	...	Fare	Cabin	Embarked
0	1	0	3	...	7.2500	No	S
1	2	1	1	...	71.2833	Yes	C
2	3	1	3	...	7.9250	No	S
3	4	1	1	...	53.1000	Yes	S
4	5	0	3	...	8.0500	No	S
5	6	0	3	...	8.4583	No	Q
6	7	0	1	...	51.8625	Yes	S
7	8	0	3	...	21.0750	No	S
8	9	1	3	...	11.1333	No	S
9	10	1	2	...	30.0708	No	C
10	11	1	3	...	16.7000	Yes	S

Figure 9 處理後欄位內容

- **One Hot Encoding**

  把 Pclass(座艙等級:1/2/3)、Age(child/young/Adult/old) 、Fare(rich /poor)、SibSp(small/middle /large)、Parch(small/middle/large) 、Cabin(Yes/No) 、Sex(male/female)做 one hot encoding

	PassengerId	Survived	Age	...	Pclass_1	Pclass_2	Pclass_3
0	1	0	22.000000	...	0	0	1
1	2	1	38.000000	...	1	0	0
2	3	1	26.000000	...	0	0	1
3	4	1	35.000000	...	1	0	0
4	5	0	35.000000	...	0	0	1
5	6	0	23.838953	...	0	0	1
6	7	0	54.000000	...	1	0	0
7	8	0	2.000000	...	0	0	1
8	9	1	27.000000	...	0	0	1
9	10	1	14.000000	...	0	1	0
10	11	1	4.000000	...	0	0	1
11	12	1	58.000000	...	1	0	0

- **Feature Selected**

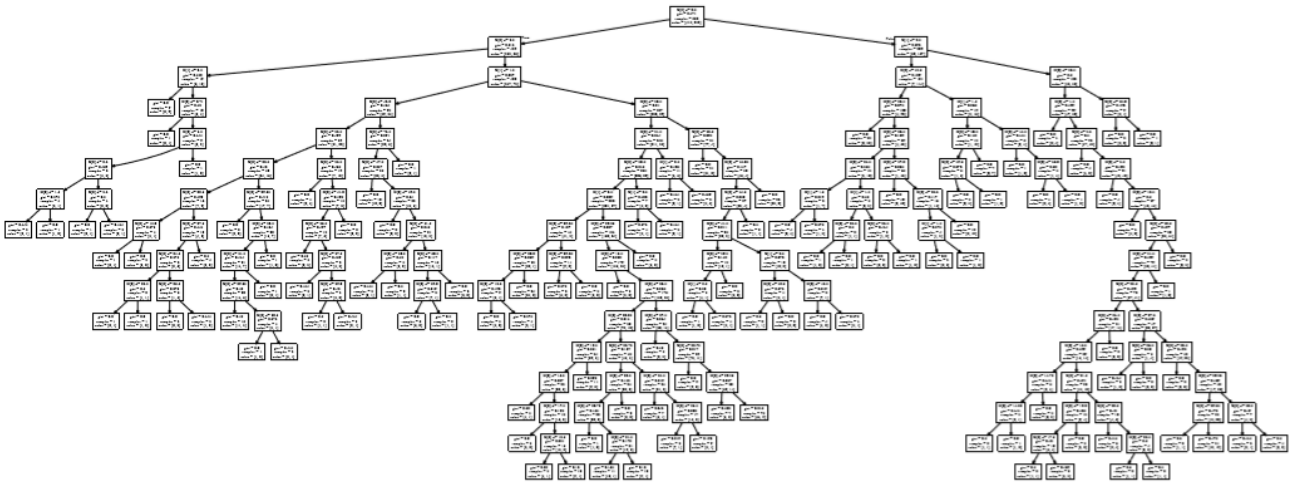
  下面我們使用此四個 feature 選法來比較分類結果。

  (a) Pclass,Age,Sex
  (b) Pclass,Age,Sex,Fare,SibSp,Parch
  (c) Pclass,Age,Sex,Fare,SibSp,Parch,Cabin,Embarked
  (d) Pclass,Age,Sex,Fare,SibSp,Parch,Cabin,Embarked(將 Age 分三個 class)

IV. Result

(a) Pclass,Age,Sex

僅選用直觀看起來最高影響力的三個 feature 來做 Decision Tree 及 Random Forest。



Age missing value 的部分用均值填滿。

Decision Tree Score :0.8340807174887892

0.8340807174887892

	precision	recall	f1-score	support
died	0.90	0.84	0.87	143
survived	0.74	0.82	0.78	80
micro avg	0.83	0.83	0.83	223
macro avg	0.82	0.83	0.82	223
weighted avg	0.84	0.83	0.84	223

Random Forest Score :0.8385650224215246

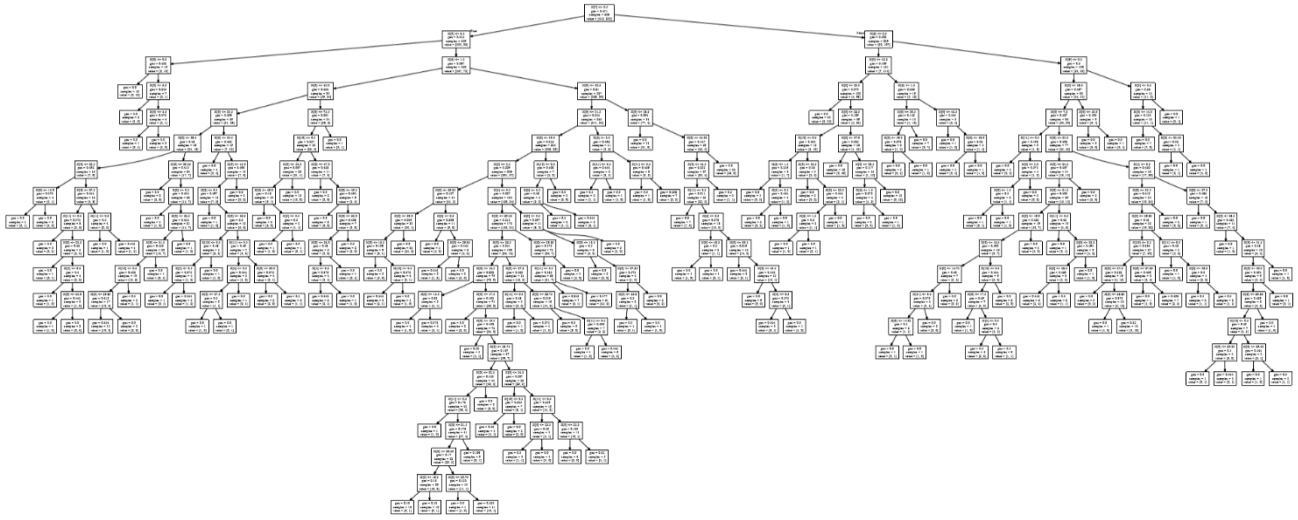
0.8385650224215246

	precision	recall	f1-score	support
0	0.89	0.85	0.87	140
1	0.76	0.82	0.79	83
micro avg	0.84	0.84	0.84	223
macro avg	0.83	0.83	0.83	223
weighted avg	0.84	0.84	0.84	223

可以看出來選用這三個 feature 就能區分出大部分的資料，有不錯的效果

(b) Pclass,Age,Sex,Fare,SibSp,Parch

接下來加入票價和親子手足個數 feature。



Decision Tree Score :0.8071748878923767

0.8071748878923767

	precision	recall	f1-score	support
died	0.87	0.82	0.84	143
survived	0.71	0.79	0.75	80
micro avg	0.81	0.81	0.81	223
macro avg	0.79	0.80	0.80	223
weighted avg	0.81	0.81	0.81	223

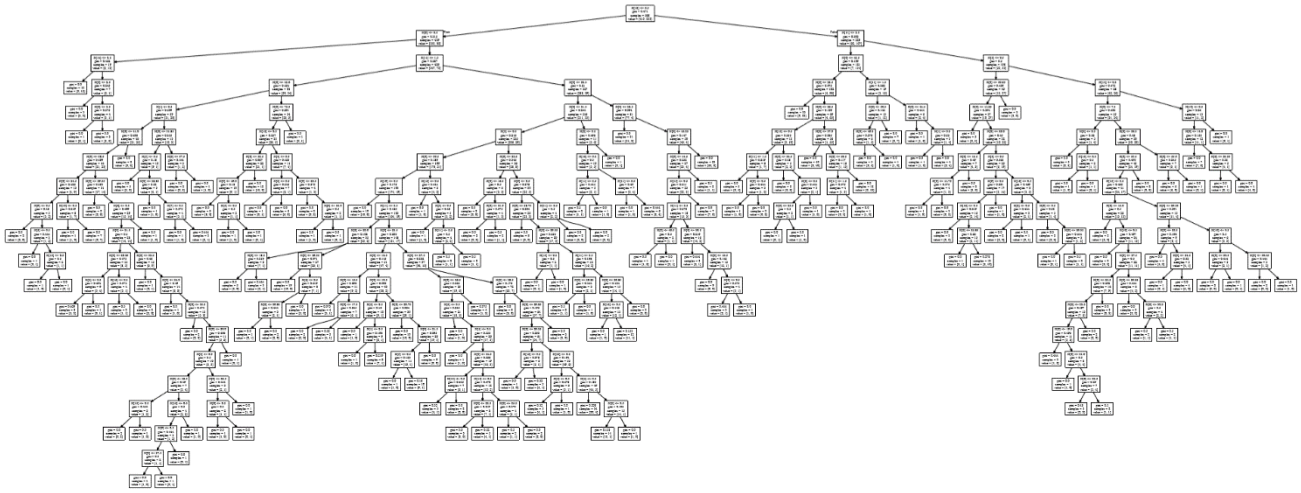
Random Forest Score :0.8161434977578476

0.8161434977578476

	precision	recall	f1-score	support
0	0.87	0.83	0.85	141
1	0.73	0.79	0.76	82
micro avg	0.82	0.82	0.82	223
macro avg	0.80	0.81	0.81	223
weighted avg	0.82	0.82	0.82	223

沒想到加入這兩個 feature 之後變差了，可能是前處理部分沒有做的很詳細，因此整個分類結果變差。

(c) Pclass,Age,Sex,Fare,SibSp,Parch,Cabin,Embarked



Decision Tree Score :0.8251121076233184

0.8251121076233184

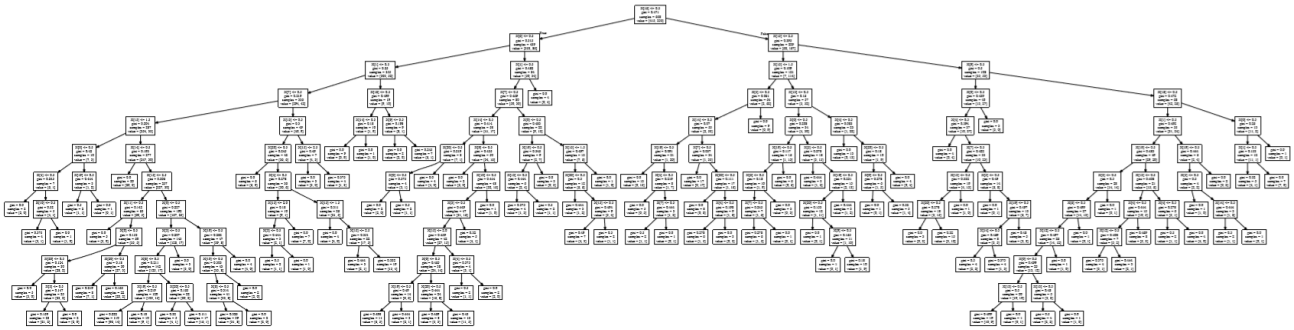
	precision	recall	f1-score	support
died	0.88	0.84	0.86	141
survived	0.74	0.80	0.77	82
micro avg	0.83	0.83	0.83	223
macro avg	0.81	0.82	0.82	223
weighted avg	0.83	0.83	0.83	223

Random Forest Score :0.8385650224215246

0.8385650224215246

	precision	recall	f1-score	support
0	0.90	0.85	0.87	142
1	0.75	0.83	0.79	81
micro avg	0.84	0.84	0.84	223
macro avg	0.82	0.84	0.83	223
weighted avg	0.84	0.84	0.84	223

(d) **Pclass,Age,Sex,Fare,SibSp,Parch,Cabin,Embarked(將 Age 分三個 class)**



這邊是將所有的 feature 做 one hot encoding 然後加入了最後兩個 feature ,以及將 Age 細分為三個 class。由 decision tree 來看深度有變小,比較 balance 一點了，結果也較為出色許多。

Decision Tee Score :0.852017937219731

```
0.852017937219731
              precision    recall  f1-score   support

        died       0.92      0.85      0.88       145
    survived       0.75      0.86      0.80        78

   micro avg       0.85      0.85      0.85       223
   macro avg       0.84      0.85      0.84       223
weighted avg       0.86      0.85      0.85       223
```

Random Forest Score :0.8654708520179372

```
0.8654708520179372
              precision    recall  f1-score   support

           0       0.90      0.88      0.89       138
           1       0.81      0.85      0.83        85

   micro avg       0.87      0.87      0.87       223
   macro avg       0.86      0.86      0.86       223
weighted avg       0.87      0.87      0.87       223
```

各 Feature Sellected 的分類結果:

Dataset	*Decision Tree*	*Random Forest*
(a)	0.834	0.838
(b)	0.807	0.816
(c)	0.825	0.838
(d)	0.865	0.865

*Decision Tree 輸出的圖存在 tree pic 資料夾中。