Data Mining

Project 2

Q36071059 蔡榮漾

本次作業採用Kaggle上的鐵達尼dataset(<https://www.kaggle.com/c/titanic/data>)

中的train.csv與test.csv，在Kaggle的Jupyter & VS Code的Python3環境上執行。

**選擇使用此資料庫的原因為：**

1. 乘客可被簡單分類為生還or罹難兩種狀態
2. 此資料庫有多種屬性可討論以及作為分類依據
3. 資料缺失數量在可接受範圍
4. **檢視資料**：

首先我先查看前10筆資料以了解本資料的格式以及屬性，並思考有那些屬性適合被拿來作為分類的依據。

除了生還與否的屬性外，其餘11個屬性代表意義如下：

PassengerId：乘客ID

Pclass：艙等

Name：乘客姓名

Sex：性别

Age：年齡

SibSp：堂兄弟/妹數

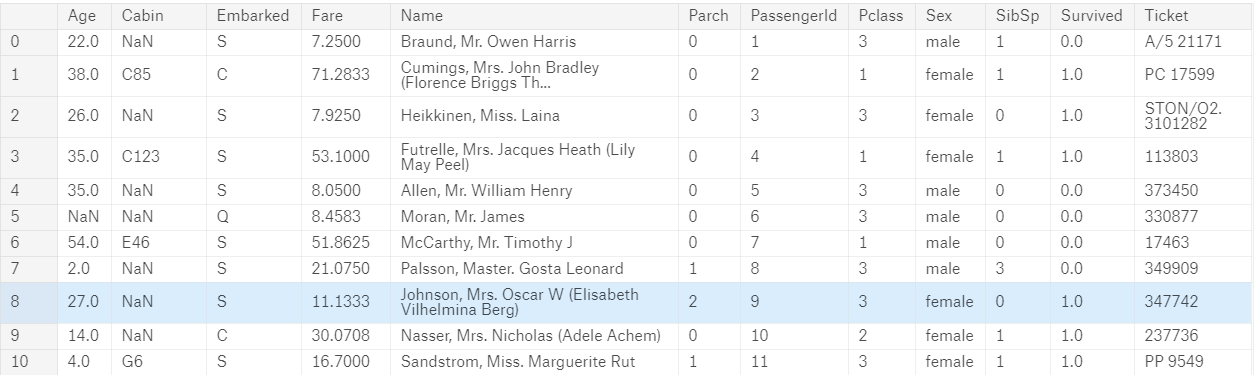
Parch：父母&小孩數

Ticket：船票資訊

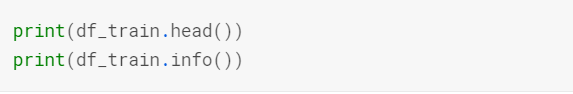
Fare ：票價

Cabin：客艙編號

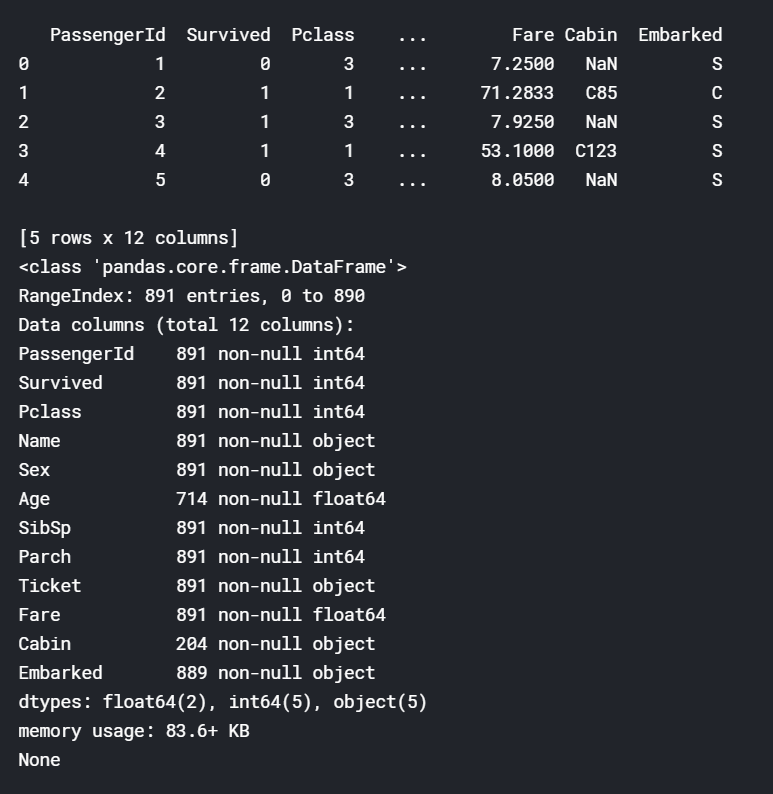
Embarked： 登船港口



接著我利用以下指令以查看資料的狀況。



得到結果：

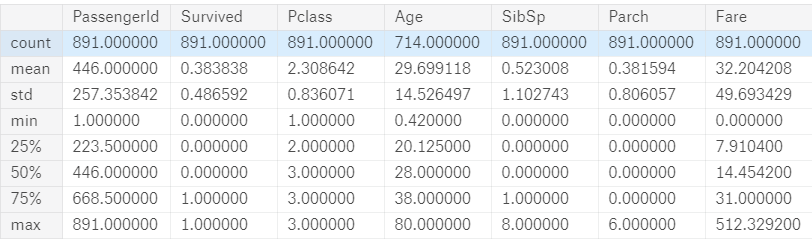


至此我們得知，Age、Cabin與Embarked有資料缺失，其中又以Cabin最為嚴重；但常理思考下，客艙編號與艙等/票價兩種屬性應會有因果上關聯，故我傾向可以直接捨棄該屬性。

而Age部分，一開始我在思考是否以平均數或中位數填滿所有缺失值，但又怕會影響分析結果，故暫時擱置。

查看各資料之分布，或許會得到有用資訊。



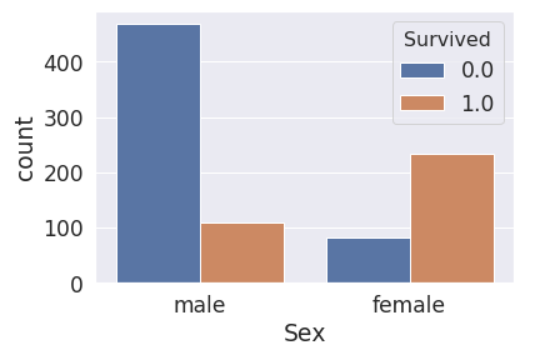
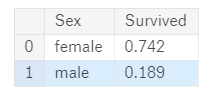


後來發現能獲得額外資訊的屬性僅有Pclass、Age、Fare三者。

考量各屬性間的特性，與衡量我個人的能力，我能夠使用的屬性應有下列四者：Pclass艙等、Sex性别、Age年齡、Fare票價 ，而艙等與票價有所掛勾，故取艙等捨棄票價。

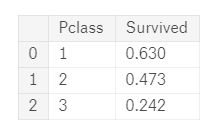
1. **屬性分析**：

首先我先查看Sex屬性與生還與否間的關係，得到：



顯然女性的生還率顯著的高於男性，推測就像電影所演，船難當下傾向讓老弱婦孺先行搭上救生艇，故資料顯示男性生還率較低也就能理解。

接著查看艙等：

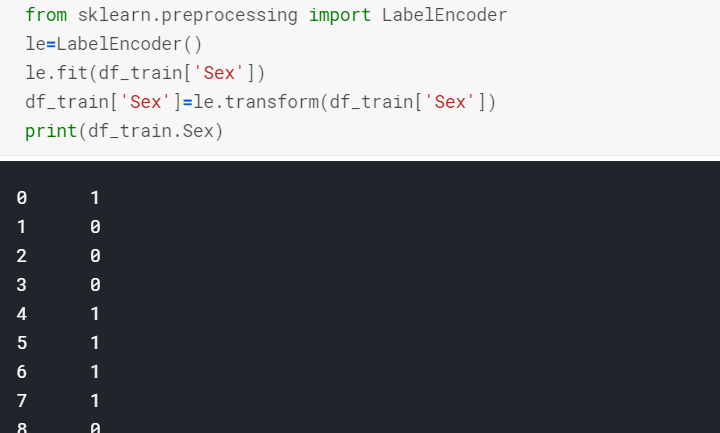


得到生還率與艙等之間的關係：艙等越高生還率越高，我一開始是推測救生艇等求生工具在較高級的艙等會較為齊全。

但依據此圖得到的資訊更像是，因為低艙等搭乘人數較多，僧多粥少的狀況下死亡人數拉高，才會造成死亡率攀升，從圖中看來三種艙等的生還人數其實是差不多的。

1. **資料處理**：

首先由於性別屬性使用的表示方式是male與female(字串)，無法直接丟進分類樹分析，先將其轉為1/0表示。



接著便是Age資料缺失部分的處理，這邊先嘗試用平均值去填滿空缺的部分，進行嘗試。

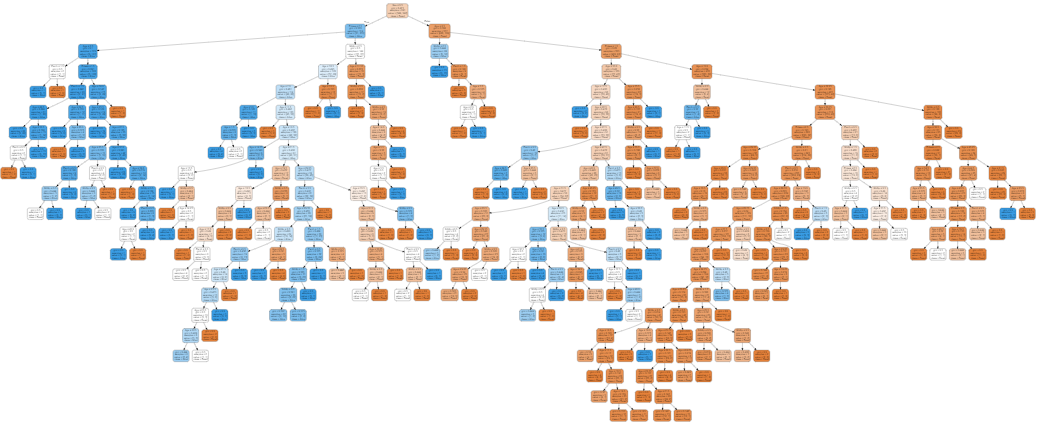


1. **決策樹分類**：

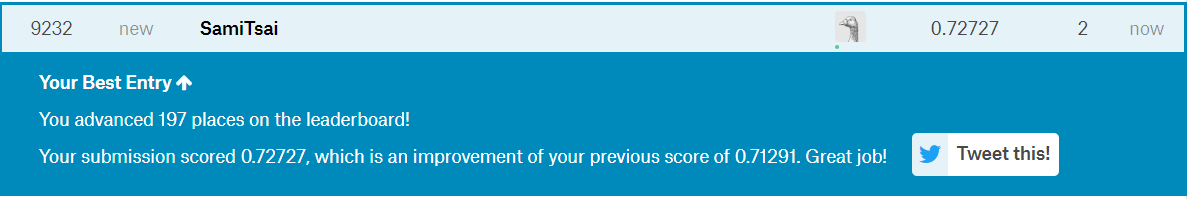
用選定的三個屬性進行決策樹的訓練，並用交叉驗證法計算此模型的準確率，得到約80.14%的準確率，並將決策樹畫出來：。







**Extra、提交Kaggle檢驗此決策樹分類準確度：**



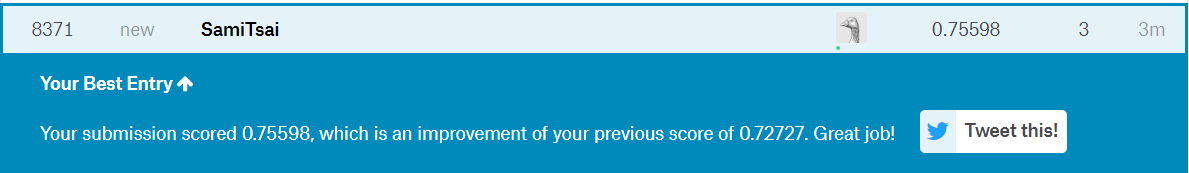
72.72%，與自行預測的80%有很大的出入，排名更是淒慘的9232/10144 XD

顯然還有很大進步空間；推測Age資料處理上使用平均數填充缺失資料的方式太為草率，以及或許可以設法充分利用其他未利用到的屬性。

**Extra2、與隨機森林比較：**

嘗試使用隨機森林進行分類



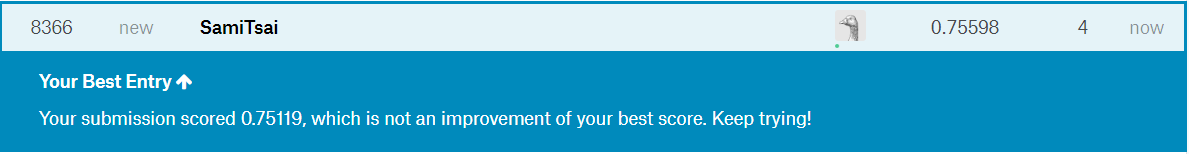
使用的特徵不變，一樣交給Kggle檢驗：

分數顯著上升至75.59%。耳聞隨機森林較能抵抗overfit，猜想或許除了Age處理手法過於粗糙外，先前決策樹的準確率偏低也有一部分是因為overfit。

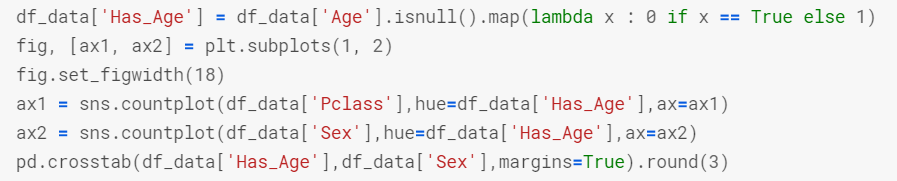
**Extra3、討論Age屬性：**

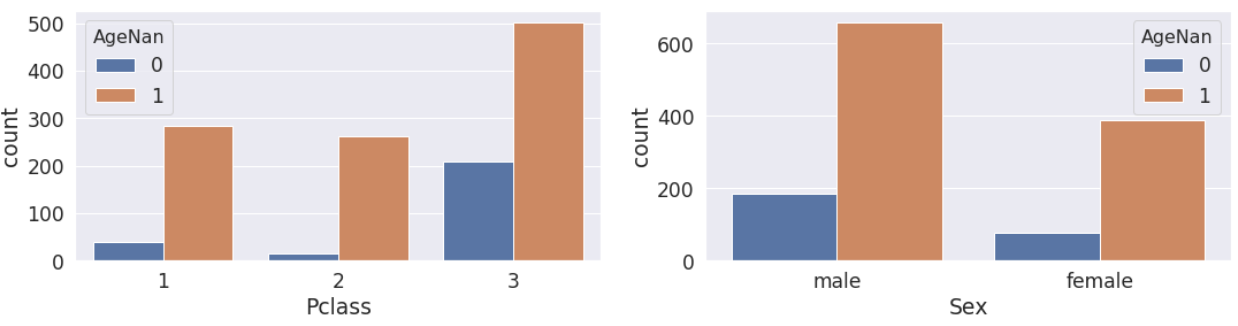
若單純使用中位數補上缺失值效果是否會比平均數好?





反而變得更差了，看來需要更進一步進行討論，先分別查看Pclass和Sex中Age屬性缺失的部分

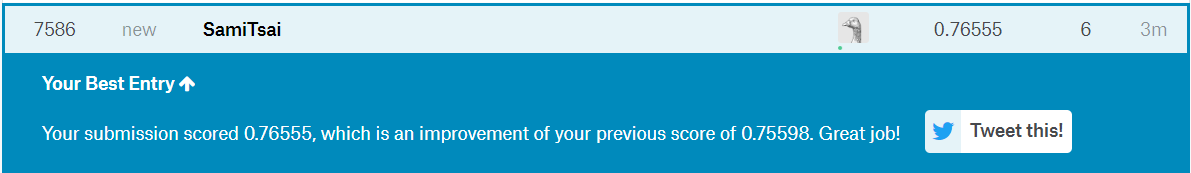




從表中得知，在最低等客艙中的乘客，缺失Age屬性的比例明顯高出其他艙等多；而男性Age屬性缺失數量較女性多，但男女在缺失比例上差不多。也就是說，年齡這個屬性不太適合用來分類整體的生還率，否則在3等艙分類上會有較嚴重的問題。

若我們排除3等艙裡的資料，拿剩下的與年齡進行分析，得到一張生還與否的圖。從圖中可以觀察到，15歲以下死亡人數趨近於零，而15歲以上則無法明顯歸納出一個規則；若我們額外創造一個屬性為”是否大於15歲”並取代Age進行分析，結果分數確實微幅的上升了。





最終排名停留在7586/10165，雖然依然不甚理想，但回頭比起最初的9232/10144依然有了顯著的進步，若能利用其他屬性創造出合理的特徵，相信能夠再繼續進步。