**111601012 大氣二 董維軒**

1. 比賽簡介: 泰坦尼克號飛船是一個月前發射的星際客輪。船上有將近13000名乘客，這艘船在前往第一個目的地時，不小心與隱藏在塵埃雲中的一個時空異常點相撞。雖然飛船完好無損，但幾乎一半的乘客都被傳送到了另一個空間！

為了幫助救援人員找回失散的乘客，你需要利用從飛船損壞的計算機系統中恢覆的記錄，預測哪些乘客被異常傳送了。

為什麼選擇這個比賽：跟上課內容比較像，對我來說比較好理解、操作、改寫，也可以從上課教的東西做發想。

資料集：train.csv:要從此資料做訓練

test.csv :以此資料驗證訓練成果

變數:

PassengerId:旅客的ID

HomePlanet:旅客出發的星球

CryoSleep: 表示乘客是否選擇在航行期間進入休眠狀態。

處於低溫睡眠狀態的乘客只能呆在自己的艙室里。

Cabin: 乘客所在艙位的編號。格式為 deck/num/side，

其中 side 可以是 P（左舷）或 S（右舷）

Destination:旅客的目的地

Age:旅客的年紀

VIP:有無購買VIP服務

RoomService, FoodCourt, ShoppingMall, Spa, VRDeck:

乘客在飛船上的設施中所花費的金額。

Name:旅客的名字

Transported: 乘客是否被傳送到了另一個空間，即預測目標。

1. 程式實作:載入train.csv和test.csv，這兩份資料都包含有關乘客的資訊，簡單分析過後，訓練模型，將test.csv由模型預測的結果存到submission.csv。

步驟如下:

(1): 載入資料：使用 Pandas 載入訓練資料和測試資料。

(2): 資料分析：簡單的分析資料分析，例如計算平均年齡，觀察各個變數與所需預測結果的關聯。

(3): 訓練模型：使用RandomForestClassifier進行模型訓練。

(4): 保存結果： 將測試資料的預測結果保存到 'submission.csv' 文件中。

(5): 寫出結果、信息：寫出模型的一些(15個)預測結果以及一條成功保存的訊息。

與上課的關聯:一些data觀察，取代資料。

延伸學習: RandomForestClassifier

參考資料:

[🚀 Spaceship Titanic: A complete guide 🏆 | Kaggle](https://www.kaggle.com/code/samuelcortinhas/spaceship-titanic-a-complete-guide/notebook)

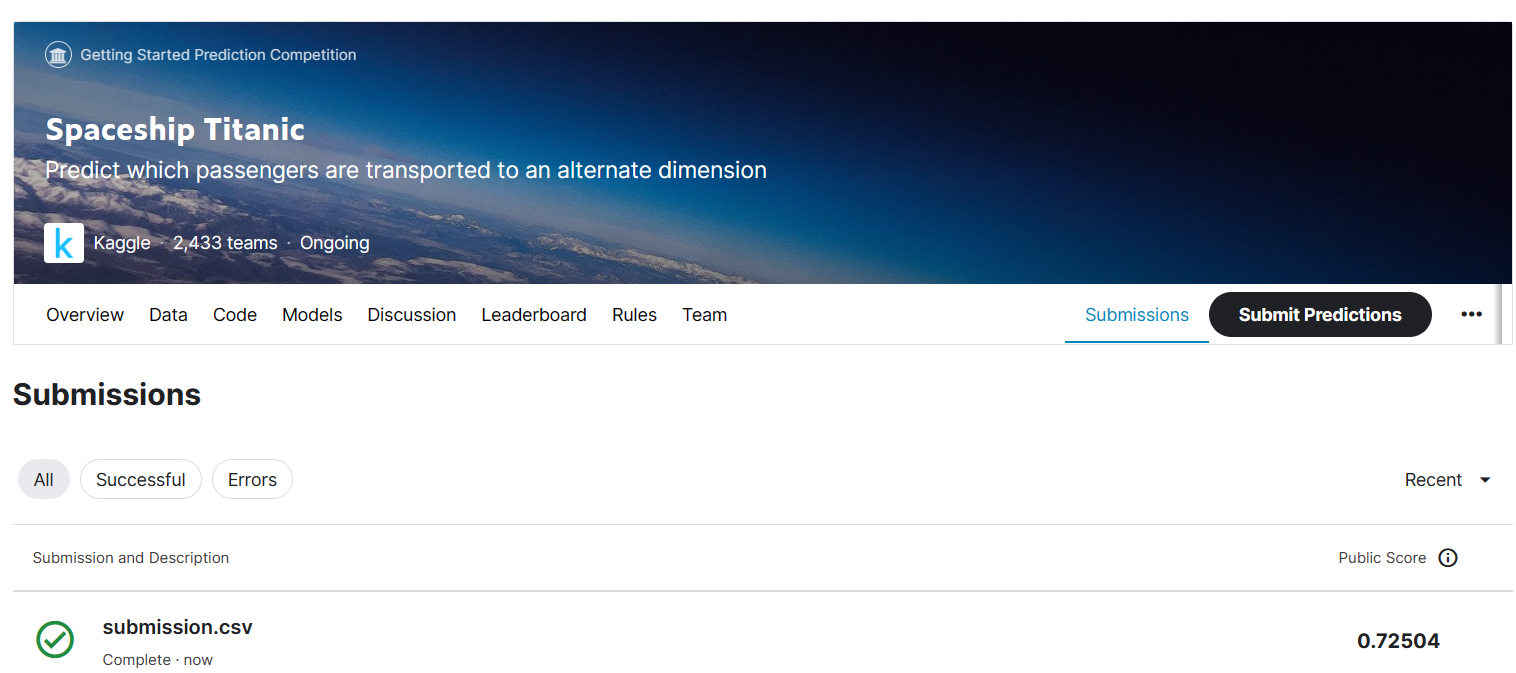
(https://www.kaggle.com/code/samuelcortinhas/spaceship-titanic-a-complete-guide/notebook)

[Pytorch 數據分析3-太空飛船泰坦尼克號預測分析 - 大大通(繁體站) (wpgdadatong.com)](https://www.wpgdadatong.com/blog/detail/70687)

(https://www.wpgdadatong.com/blog/detail/70687)

[How To Score ~80% Accuracy in Kaggle’s Spaceship Titanic Competition | by Zaynab Awofeso | CodeX | Medium](https://medium.com/codex/how-to-score-80-accuracy-in-kaggles-spaceship-titanic-competition-using-random-forest-classifier-e7d06ce25bad)

(https://medium.com/codex/how-to-score-80-accuracy-in-kaggles-spaceship-titanic-competition-using-random-forest-classifier-e7d06ce25bad)

1. 結果: 

改善:有些變數沒有使用到:

Cabin, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck，將全部的變數都用上應該才會是最準的預測。

還有取代空白資料用的方式是用最多的資料直接填入，理想上應該是要用已知的比例去分配。

以及取代的部分有一些會出Bug，不清楚原因，暫時用#來無視。

目前用的方法是簡單分析，各個變數彼此獨立(單變數分析)，未考慮到互相影響的情況(多變數分析)

1. Code:

