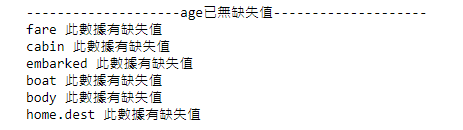
HW10 B063012054 林祐安

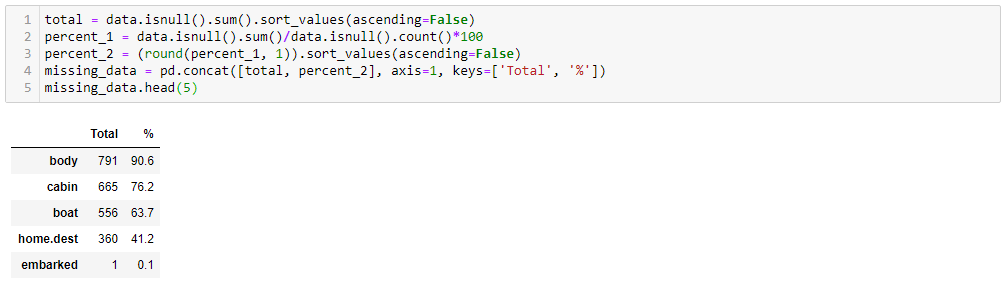
Data Transform

1. 第一個處理的為age，直接套用助教的方法



1. 根據結果，依序處理下一個為fare，就我們的常識而言年紀小或大通常很便宜，其他就成人價，在此不考慮身心障礙等，因此我將年紀區分成小於12歲、12歲以上小於65歲、65歲以上，三種狀況，將三個區間取票價平均填入缺失值。

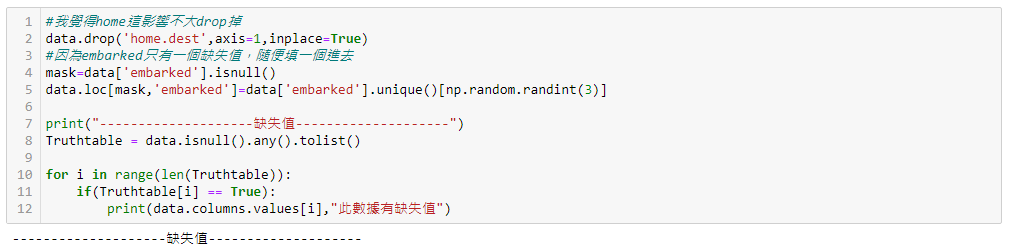


1. 下一組為cabin，觀察了cabin數據發現是代號，因此使用

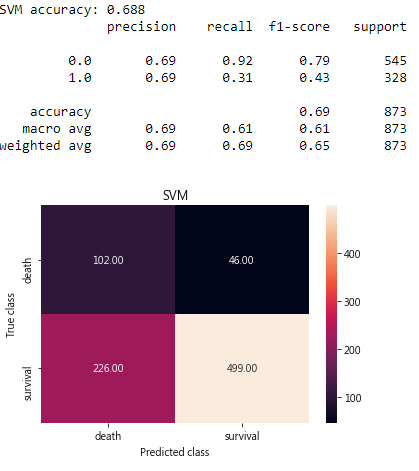
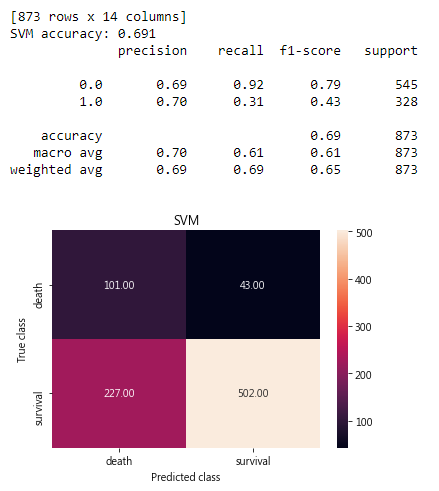
查看缺失值占比，發現body、cabin這兩組缺失太多，而cabin還是代號，因此drop掉，不浪費時間處理，隨意填值並不是好方法，缺失太多資料寧可丟棄。觀察body可以發現，其值出現在沒有搭上船的，可以認為這是屍體，因此body與boat應有相關性，我取boat當代表，沒有搭上船就當作死了，搭上船當作活的，因此可以drop掉body。



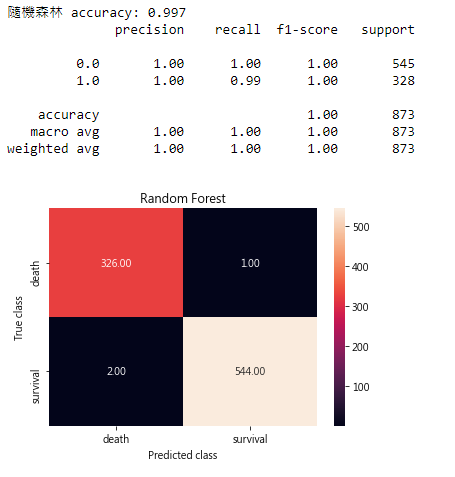
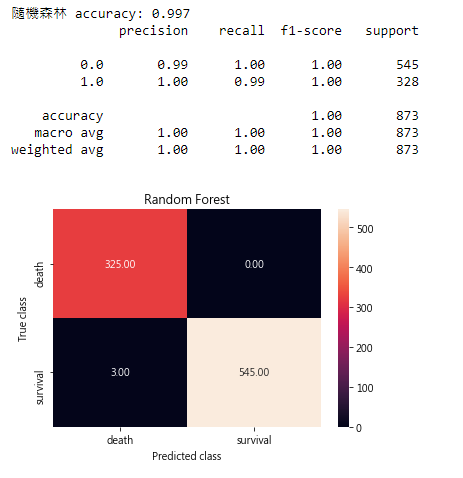
1. 剩下embarked與home.dest，其中home.dest是地名，就常理而言，死人不會有目的地，但實際上觀察發現，死人也有值，因此可以推論是出發前決定的目的地，或是活著到、死了被打撈到的地點，很明顯，這資料跟我們要預測會不會活命是沒有關聯的，因此drop掉。
2. 最後是embarked，僅缺失一個值，因此隨便填一個進去，影響不大



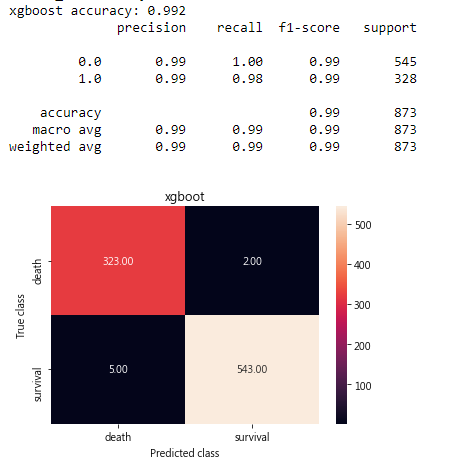
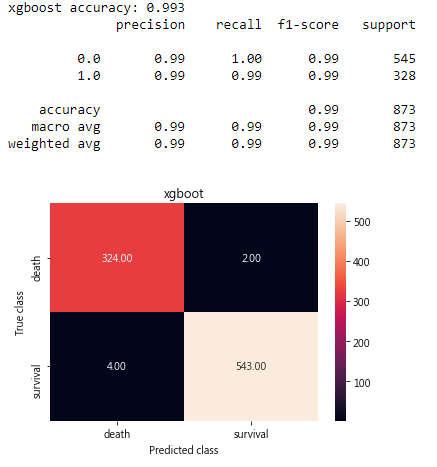
1. 預測結果



(我的) (助教的)



(我的) (助教的)



(我的) (助教的)

由於有刪掉一些資訊，可以預期正確率應該會稍微下降一些，在SVM中比較明顯，SVM的演算法比起其他兩個模型較單純，因此直接刪除值的行為影響決策幅度會比較大，另外兩個模型演算法較複雜，考慮的面相較多，其結果值若我們刪掉一些比較不重要的資訊，理應變化不大，當資料龐大時，還是得先過濾一些常理上不需要的資訊，減少資料量，缺失值若太多，選擇淘汰掉應是比較好的選擇，否則這份資料就不像是根據實際狀況統計得結果，而是藉由一些演算法而生成的資料。

最後將檔案輸出

