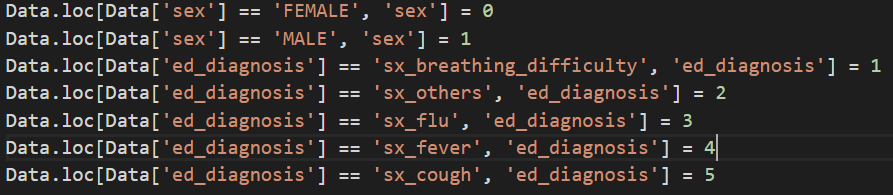
Assignment2 report 107062338 邱俊維

一、Data preprocessing

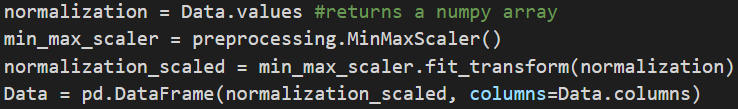
這次的testing data中missing的部分會用眾數來補上，因此在training的部分為貼近實際方式，也採用了補上眾數的策略。



Data有連續以及非連續的的資料，其中” ed\_diagnosis”欄位有五種不同的種類，因此將其以1~5取代，”sex”欄位也使用同樣的方法。

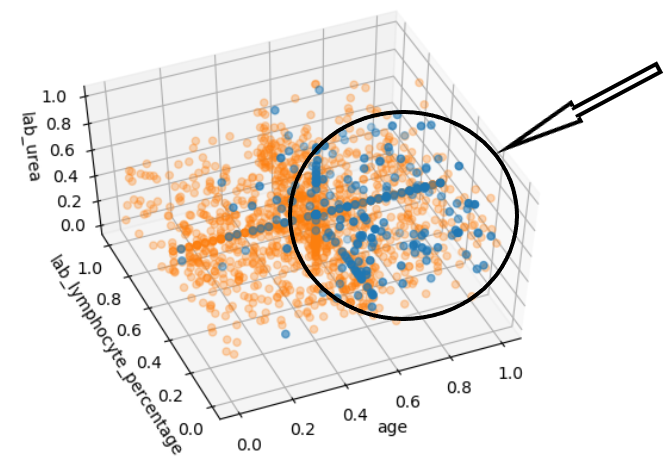


將連續的資料進行正則化(normalization)。



為了更有效的分析各項資料之間的相互關係，以及與最後的結果的相關程度，我嘗試將各種Attributes組合描繪成3D圖形。

Ex: age、lab\_lymphocyte\_percentage、lab\_urea

 (藍點為死亡)

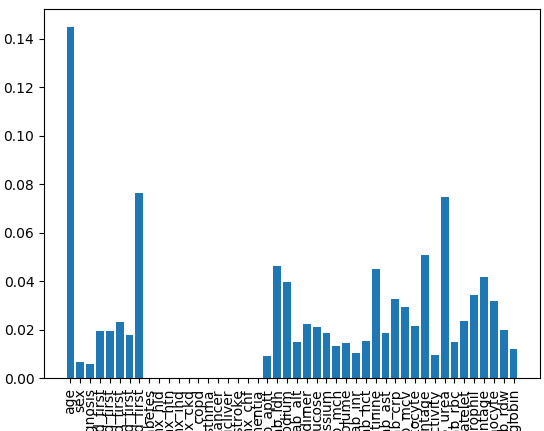
從上圖可以看出藍點分布在age較高的地方，且相關性頗高。而另外兩個較弱的資訊lab\_urea、lab\_lymphocyte\_percentage。死亡大致分布在lab\_urea較高的地方，而分布在lab\_lymphocyte\_percentage較小的地方。藉由此方法來挑選適合使用的Attributes。

二、Models

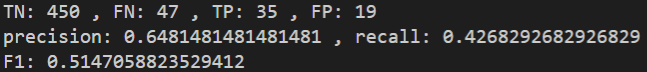
嘗試使用了各式介紹過的classifier，例如：SVM、Decision Tree、NN、Random Forest等。SVM的效果糟糕，Decision Tree的precision過低，而NN雖然precision可達0.8，但也因此recall大約只有30。最後選擇使用precision以及recall表現較平均的Random Forest作為預測的model.

原本只使用經過分析後覺得與結果較有相關性的features作為input data，但在後來發現若將所有features都當作input data進行training，效果較好一些。

Random Forest model建起來後，觀察data中的各個attributes的重要性，並將其繪製成長條圖。可見如同先前對資料進行觀察時所發現，age是一項重要的依據，而另外還有數樣比重較高的如vitals\_spo2\_ed\_first、lab\_urea等。



Data以7:3的比例分開，7成作為training data，3成作為testing data。在這樣的情境下，precision及recall約為各五成上下左右。



為避免model的overfitting，在random forest model中加入了minimum samples leaf = 7的限制。



How to use the model file、

File name:107062338\_HW2\_Model.py

Required file: hm\_hospitales\_covid\_structured\_30d\_train.csv、split\_train\_export\_30d.csv、fixed\_test.csv

執行107062338\_HW2\_Model.py後會output出107062338.csv，即為預測結果。