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 = Data.fillna(Data.mode().iloc[0])
```

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。

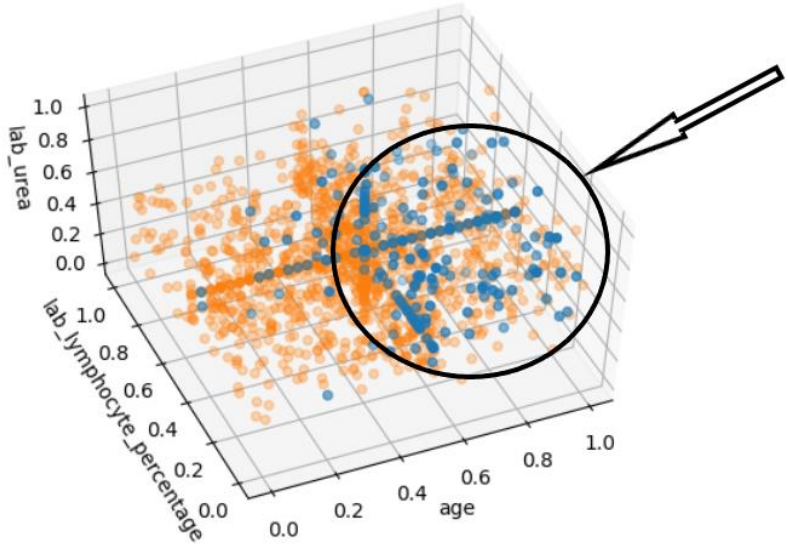
```
Data.loc[Data['sex'] == 'FEMALE', 'sex'] = 0
Data.loc[Data['sex'] == 'MALE', 'sex'] = 1
Data.loc[Data['ed_diagnosis'] == 'sx_breathing_difficulty', 'ed_diagnosis'] = 1
Data.loc[Data['ed_diagnosis'] == 'sx_others', 'ed_diagnosis'] = 2
Data.loc[Data['ed_diagnosis'] == 'sx_flu', 'ed_diagnosis'] = 3
Data.loc[Data['ed_diagnosis'] == 'sx_fever', 'ed_diagnosis'] = 4
Data.loc[Data['ed_diagnosis'] == 'sx_cough', 'ed_diagnosis'] = 5
```

將連續的資料進行正則化(normalization)。

```
normalization = Data.values #returns a numpy array
min_max_scaler = preprocessing.MinMaxScaler()
normalization_scaled = min_max_scaler.fit_transform(normalization)
Data = pd.DataFrame(normalization_scaled, columns=Data.columns)
```

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)

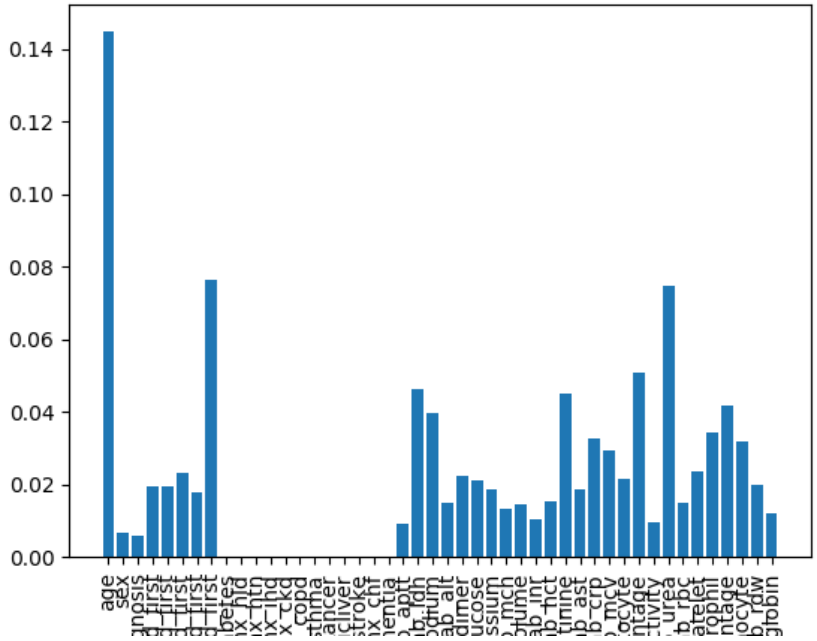
從上圖可以看出藍點分布在 age 較高的地方，且相關性頗高。而另外兩個較弱的資訊 lab_urea、lab_lymphocyte_percentage。死亡大致分布在 lab_urea 較高的地方，而分布在 lab_lymphocyte_percentage 較小的地方。藉由此方法來挑選適合使用的 Attributes。

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。

```
TN: 450 , FN: 47 , TP: 35 , FP: 19
precision: 0.6481481481481481 , recall: 0.4268292682926829
F1: 0.5147058823529412
```

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

```
clf = RandomForestClassifier(min_samples_leaf= 7)
```

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。