Final-Team11 Arteon

陳泓吟 工資 112 H34086254

陳宜姍 統計 114 C14101137

陳凱騫 統計 114 H24101222

一.主題及問題描述

我們的主題為新冠肺炎確診病患的死亡預測。利用確診新冠肺炎病患的各項醫療指標預測其死亡的可能性，該技術應用的預測結果能協助醫護人員加強關注高風險確診者。

二.輸入/輸出

該主題使用的資料存在 kaggle，來源為墨西哥政府提供的醫療數據。原始資料總共有 21 項欄位，1048576 筆資料。21 項欄位說明如表(一)所示。

欄位名稱	欄位說明	資料類型
age	age of the patients	連續型資料
sex	1 for female and 2 for male	類別型資料
classification	covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive	類別型資料
patient type	type of care the patient received in the unit. 1 for returned home and 2 for hospitalization	類別型資料

usmr	Indicates whether the patient treated medical units of the first, second or third level	類別型資料
medical unit	type of institution of the National Health System that provided the care	類別型資料
pneumonia	whether the patient already have air sacs inflammation or not	布林值
pregnancy	whether the patient is pregnant or not	布林值
diabetes	whether the patient has diabetes or not	布林值
copd	Indicates whether the patient has Chronic obstructive pulmonary disease or not	布林值
asthma	whether the patient has asthma or not	布林值
inmsupr	whether the patient is immunosuppressed or not	布林值
hypertension	whether the patient has hypertension or not	布林值
cardiovascular	whether the patient has heart or blood vessels related disease	布林值
renal chronic	whether the patient has chronic renal disease or not	布林值
other disease	whether the patient has other disease or not	布林值

obesity	whether the patient is obese or not	布林值
tobacco	whether the patient is a tobacco user	布林值
intubed	whether the patient was connected to the ventilator	布林值
icu	Indicates whether the patient had been admitted to an Intensive Care Unit	布林值
date died	If the patient died indicate the date of death, and 9999-99-99 otherwise	日期

表(一)欄位說明

21 項欄位中包含連續型資料欄位、類別型資料欄位及布林值資料欄位。布林值資料欄位以 1 表示正面資料，2 表示負面資料，而 97、99 為無效值，需另外處理。

其中 21 項欄位中 20 項為原始資料的訓練輸入值，剩餘一欄位 date died 則須自行將其轉換為目標欄位。date died 欄位紀錄病患死亡日期，若該病患存活則會以 9999-99-99 表示。因此我們會將 date died 欄位轉為 death 欄位，同樣以 1、2 表示，1 表示死亡，2 表示存活。轉換方式如圖(一)。並且之後將 date died 欄位刪除。

```
# If we have "9999-99-99" values that means this patient is alive.
df["DEATH"] = [2 if each=="9999-99-99" else 1 for each in df.DATE_DIED]
```

圖(一)目標欄位轉換

三.挑戰

該題目的挑戰有三。第一為需自行轉換目標欄位，第二為特徵選擇，第三為目標欄位的不平衡資料處理。自行轉換目標欄位部分已如上三.所說明成功進行轉換，其餘兩項挑戰則在以下篇幅說明如何進行處理。

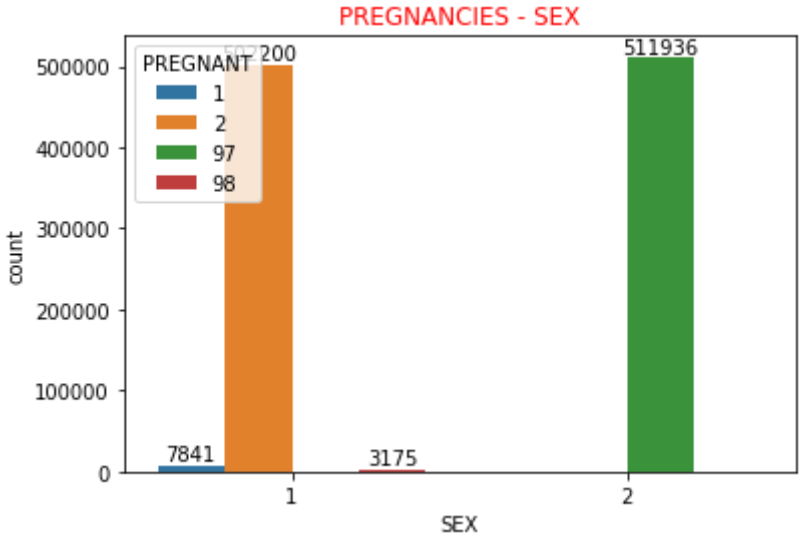
四.特徵處理與分析

針對布林值資料欄位，除了 intubed、pregnant 及 icu 欄位外，我們將其他布林值欄位值為 97、98 無效值的資料列直接進行刪除只保留有效值。如圖(二)所示。

```
df = df[(df.PNEUMONIA == 1) | (df.PNEUMONIA == 2)]
df = df[(df.DIABETES == 1) | (df.DIABETES == 2)]
df = df[(df.COPD == 1) | (df.COPD == 2)]
df = df[(df.ASTHMA == 1) | (df.ASTHMA == 2)]
df = df[(df.INMSUPR == 1) | (df.INMSUPR == 2)]
df = df[(df.HIPERTENSION == 1) | (df.HIPERTENSION == 2)]
df = df[(df.OTHER_DISEASE == 1) | (df.OTHER_DISEASE == 2)]
df = df[(df.CARDIOVASCULAR == 1) | (df.CARDIOVASCULAR == 2)]
df = df[(df.OBESITY == 1) | (df.OBESITY == 2)]
df = df[(df.RENAL_CHRONIC == 1) | (df.RENAL_CHRONIC == 2)]
df = df[(df.TOBACCO == 1) | (df.TOBACCO == 2)]
```

圖(二)保留有效資料

接著 pregnant 欄位需按照性別分別進行處理。將懷孕與否和性別進行比對，結果如圖(三)所示。



圖(三) pregnant-sex 欄位比對

男性部分，懷孕欄位應皆為否定值，因此我們將 97 皆轉換為 2，符合否定欄位的表示方式。而女性部分，僅保留欄位值為 1、2 的有效資料。轉換方式如圖(四)所示。

```
# Converting process according to inference above
df.PREGNANT = df.PREGNANT.replace(97,2)

# Getting rid of the missing values
df = df[(df.PREGNANT == 1) | (df.PREGNANT == 2)]
```

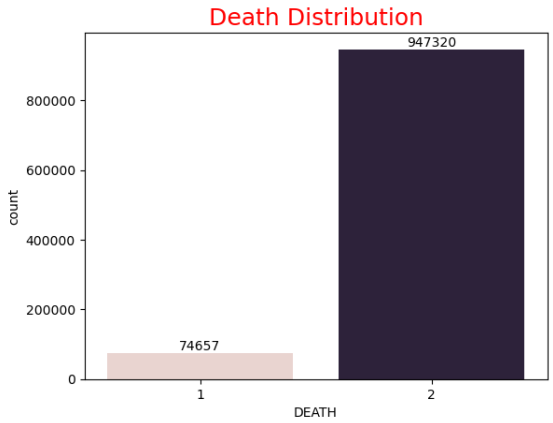
圖(四)pregnant 欄位處理

其餘兩項布林值欄位 intubed、icu 因有 830385 筆資料皆為缺失值 97，因此我們直接刪除 intubed、icu 兩欄位。

將所有缺失值進行轉換刪除等方式處理過後，餘 1021977 筆資料及 18 項輸入欄位。

五.資料觀察

目標欄位 death 的分布圖如下圖(五)，由圖可看出其分布極為不平衡，因此必須經過 resampling 處理該不平衡的問題，而我們選用的是 Nearmiss 方法，處理方法及結果如圖(六)。



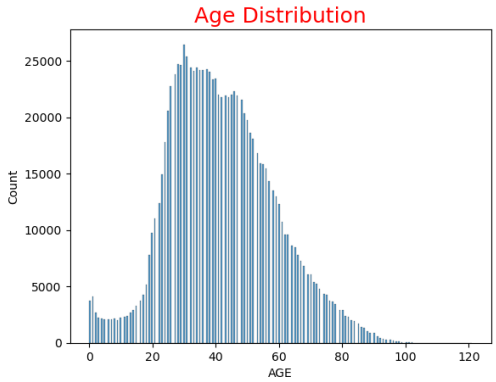
圖(五) 目標欄位分布

```
from collections import Counter
from imblearn.under_sampling import NearMiss
print('Original dataset shape %s' % Counter(train_y))
nm = NearMiss()
X_res, y_res = nm.fit_resample(train_x, train_y)
print('Resampled dataset shape %s' % Counter(y_res))
```

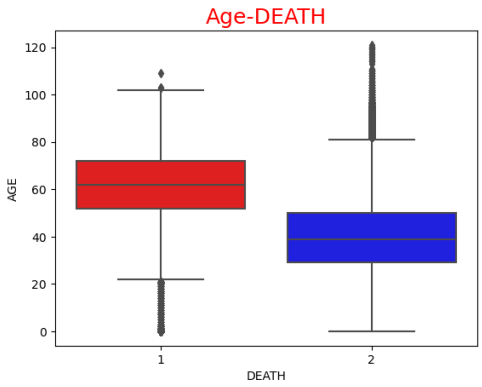
```
Original dataset shape Counter({2: 757723, 1: 59858})
Resampled dataset shape Counter({1: 59858, 2: 59858})
```

圖(六) 目標欄位正反比例不平衡處理

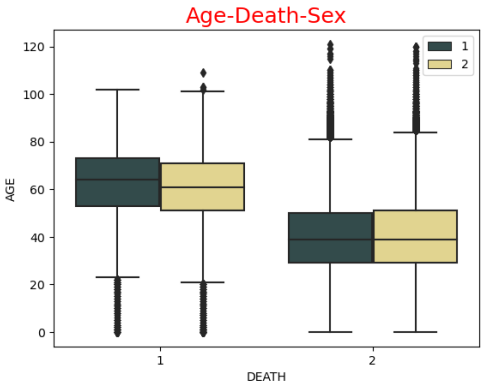
其他的資料觀察還包括確診病患的年齡分布、年齡性別及死亡的關係比對，分別如圖(七)、(八)、(九)所示。由圖可看出年齡較大的病患其死亡的風險較高。而性別與是否死亡則沒有太大關聯。



圖(七)確診病患年齡分布



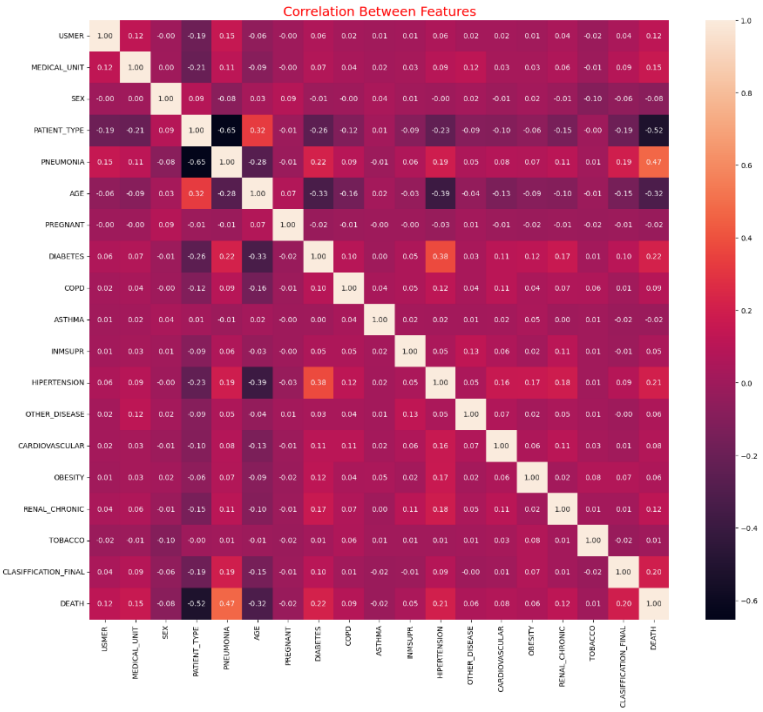
圖(八)死亡與否的年齡分布圖



圖(九)死亡與否的年齡及性別分布

六. 資料前處裡

最後再進入模型訓練前，我們透過 heatmap 觀察各欄位與目標欄位的相關性，如圖(十)所示。並且將與目標欄位低度相關的欄位去除，包括"SEX"、"PREGNANT"、"COPD"、"ASTHMA"、"INMSUPR"、"OTHER_DISEASE"、"CARDIOVASCULAR"、"OBESITY"、"TOBACCO"，刪除後共剩 9 個欄位。



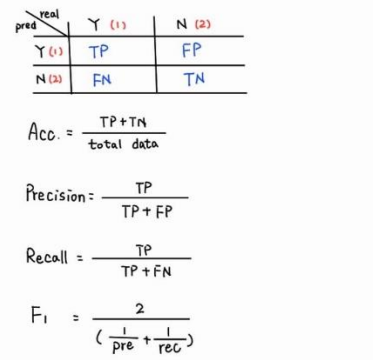
圖(十) 欄位間相關性圖

剩餘欄位中，我們針對非 binary 的類別型欄位進行 dummy，dummy 過後的欄位是以 0 表示負面資料，1 表示正面資料，我們將其轉換為與其他欄位相同以 1 表示正面資料、2 表示負面資料。另外連續型資料欄位 age 則透過 StandardScaler 進行標準化。

資料全數處理完畢後剩餘 1021977 筆資料，將其切割為 0.8 比例的訓練資料及 0.2 比例的測試資料，即可開始進行模型訓練及預測。

七. 模型評估方式

我們的模型是使用 accuracy、precision、recall 及 f1-score 進行準確率評估。評估計算方式如下圖(十一)。在資料正反比例不平衡的情形下使用 accuracy 會比較不具參考價值，而 precision 和 recall 則較著重在被預測為 positive 的資料，f1-score 則為 precision 及 recall 的綜合指標。若以本題目來說，可能會較希望著重在死亡的 recall 值上，期望可準確找出所有最終為死亡的病患。



圖(十一)模型評估方法

八.模型訓練

我們選用的模型有 LogisticRegression、DecisionTreeClassifier、BaggingClassifier、ExtraTreesClassifier、HistGradientBoostingClassifier 及 VotingClassifier 等不同模型。另外我們會以同一組模型參數是否經過 NearMiss 處理不平衡資料進行前後比較(以下模型結果第一張為未經過 NearMiss 處理，第二張為有經過 NearMiss 處理)，觀察其對不同準確率指標的影響。

LogisticRegression

透過 cross_val_score 調整模型參數，測試的參數範圍如下表：

C	[0.001,0.01,0.1,1,10]

最後選用的模型參數為 LogisticRegression(C=10)及模型結果如下:

```
Logistic Regression Accuracy : 0.9393676979980039
              precision    recall  f1-score   support

           1       0.61      0.45      0.52     14799
           2       0.96      0.98      0.97    189597

    accuracy                           0.94    204396
   macro avg       0.78      0.71      0.74    204396
weighted avg       0.93      0.94      0.94    204396
```

```
Logistic Regression Accuracy : 0.8766658838724828
              precision    recall  f1-score   support

           1       0.35      0.79      0.48     14799
           2       0.98      0.88      0.93    189597

    accuracy                           0.88    204396
   macro avg       0.66      0.84      0.71    204396
weighted avg       0.94      0.88      0.90    204396
```

DecisionTreeClassifier

透過 cross_val_score 調整模型參數，測試的參數範圍如下表：

max_depth	[5,6,7,8]

最後選用的模型參數為

tree.DecisionTreeClassifier(criterion= 'entropy',max_depth= 8)

其模型結果如下:

Decisiontree Tree Accuracy : 0.9396367834986986

	precision	recall	f1-score	support
1	0.62	0.42	0.50	14799
2	0.96	0.98	0.97	189597
accuracy			0.94	204396
macro avg	0.79	0.70	0.74	204396
weighted avg	0.93	0.94	0.93	204396

Decisiontree Tree Accuracy : 0.6287892130961467

	precision	recall	f1-score	support
1	0.14	0.81	0.24	14799
2	0.98	0.61	0.75	189597
accuracy			0.63	204396
macro avg	0.56	0.71	0.50	204396
weighted avg	0.92	0.63	0.72	204396

BaggingClassifier+ DecisionTreeClassifier

透過 cross_val_score 調整 BaggingClassifier 模型參數，測試的參數範圍如下表：

n_estimators	[50,100,150]

最後選用的模型參數為 BaggingClassifier(base_estimator= clf,n_estimators= 100, n_jobs= -1)，其模型結果如下:

decisiontree+bagging Accuracy : 0.9401455997181941

	precision	recall	f1-score	support
1	0.63	0.43	0.51	14799
2	0.96	0.98	0.97	189597
accuracy			0.94	204396
macro avg	0.79	0.71	0.74	204396
weighted avg	0.93	0.94	0.93	204396

```
decisiontree+bagging Accuracy : 0.6236961584375428
              precision    recall  f1-score   support

           1       0.15      0.87      0.25     14799
           2       0.98      0.60      0.75    189597

    accuracy                           0.62    204396
   macro avg       0.56      0.74      0.50    204396
weighted avg       0.92      0.62      0.71    204396
```

ExtraTreesClassifier

透過 cross_val_score 調整模型參數，測試的參數範圍如下表：

n_estimators	[50,100,150]
max_depth	[5,6,7,8]

最後選用的模型參數為 ExtraTreesClassifier(n_estimators=150,max_depth=8)，其模型結果如下:

```
ExtraTreesClassifier: 0.9381201197675101
              precision    recall  f1-score   support

           1       0.67      0.28      0.40     14799
           2       0.95      0.99      0.97    189597

    accuracy                           0.94    204396
   macro avg       0.81      0.64      0.68    204396
weighted avg       0.93      0.94      0.93    204396
```

```
ExtraTreesClassifier: 0.7820554218282159
              precision    recall  f1-score   support

           1       0.23      0.87      0.37     14799
           2       0.99      0.78      0.87    189597

    accuracy                           0.78    204396
   macro avg       0.61      0.82      0.62    204396
weighted avg       0.93      0.78      0.83    204396
```

HistGradientBoostingClassifier

HistGradientBoostingClassifier 為一個適用大量數據的模型，該模型我們使用預設的參數。其模型結果如下:

```
HistGradientBoostingClassifier : 0.9415399518581576
              precision    recall  f1-score   support

           1       0.63      0.46      0.53     14799
           2       0.96      0.98      0.97    189597

    accuracy                           0.94    204396
   macro avg       0.80      0.72      0.75    204396
weighted avg       0.94      0.94      0.94    204396
```

HistGradientBoostingClassifier : 0.42802207479598425

	precision	recall	f1-score	support
1	0.10	0.84	0.18	14799
2	0.97	0.40	0.56	189597
accuracy			0.43	204396
macro avg	0.53	0.62	0.37	204396
weighted avg	0.91	0.43	0.53	204396

九.實驗結果及分析

根據以上各個模型的測試結果均可以發現一個趨勢，當透過 NearMiss 處理目標欄位的不平衡後，可有效提升死亡的 recall 值，但相對 precision 即會降低。但如上述模型評估處所提，我們主要希望能提升死亡的 recall 值，準確找出最後實際情況為死亡的病患。因此死亡 recall 值的提升是我們所樂見的。然而因為 DecisionTreeClassifier、HistGradientBoostingClassifier 等數個模型在提升死亡 recall 值的同時，precision 的降低幅度過大，若我們能將 recall 提升即 precision 降低達到折衷平衡或許也是一個不錯的選擇。在以上測試的模型中 LogisticRegression 能達到最好的折衷效果，另外我們也嘗試使用 VotingClassifier 結合不同的模型進行預測，試圖將兩個指標達到折衷效果。我們嘗試了兩種 VotingClassifier 的模型，分別是 LogisticRegression 結合 HistGradientBoostingClassifier 及 LogisticRegression 結合 BaggingClassifier 得到的模型結果如下兩圖。但數據顯示這兩個模型的結果仍沒有 LogisticRegression 佳，因此判斷在該題目中， 若想將死亡 recall 值提升而又不想讓 precision 下降太多 LogisticRegression 會是最佳的選擇。

VotingClassifier : 0.613612790856964

	precision	recall	f1-score	support
1	0.14	0.83	0.24	14799
2	0.98	0.60	0.74	189597
accuracy			0.61	204396
macro avg	0.56	0.71	0.49	204396
weighted avg	0.92	0.61	0.70	204396

VotingClassifier : 0.6783987944969568

	precision	recall	f1-score	support
1	0.16	0.84	0.27	14799
2	0.98	0.67	0.79	189597
accuracy			0.68	204396
macro avg	0.57	0.75	0.53	204396
weighted avg	0.92	0.68	0.76	204396

十.結論

這次選擇的題目讓我們學習到更多處理資料及模型預測的方法，而該這些法除了應用在新冠肺炎預測外，也能同樣運用到其他重大疾病的死亡風險預測，作為醫療服務上的輔助。