第六組\_DM\_hw02

組員:R07725049吳姿君、R08725008周若涓、R08725010陳亦珊、R08725030徐薇尹

**一、研究動機與目的**

因應影響全球的疫情，觀察到疫情的分佈受到年齡、居住地、旅遊史等參數影響，除此之外，我們更進一步分析，再加上症狀的分析(如: pneumonia, fever, cough)，是否可以預測感染者是否能康復出院抑或是導致死亡的結果。

**二、資料描述**

資料來源: <https://github.com/beoutbreakprepared/nCoV2019/blob/master/latest_data/latestdata.csv>

原始資料包含以下欄位:

age, sex, city, province, country, symptoms, lives\_in\_Wuhan, travel\_history\_location, additional\_information, chronic\_disease\_binary, chronic\_disease, outcome

我們將資料做了以下調整與標示:

* 空值填入 unknown
* outcome 欄位的內容過多分類值，如: death, recovered, recovering at home, released from quarantine, severe等；因此，因此，以 outcome 的內容為基礎，判斷狀態是否 alive，新增 alive 欄位標示 yes/no
* 將 age 欄位以10為單位分為區間，每個數值以其區間之中位數做為表示，形成新的欄位 age\_range，例如: 0-9，標 5
* age 欄位的空值標 55
* age 欄位range 太大的原始資料(e.g. 10-80)的就視為空值，標 55
* symptoms 欄位的資料，原始資料是以文字描述，我們文字段落資料，切割形成不同症狀 label 欄位(如: pneumonia, fever, cough)， 最終大約將 symptoms 欄位拆成 51 個症狀欄位，原始資料有描述到的症狀標示 1，反之，標示 0
* symptoms 欄位的資料為 unknown(空值) 與 mild 就把症狀全部標 0

經過前處理後共留下 1467 筆資料

調整後的資料包含以下欄位: age\_range, sex, city, province, country, systemic weakness, somnolence, obnubilation, primary myelofibrosis, multiple electrolyte imbalance, multiple organ failure, myocardial infarction lesions on chest radiographs, eye irritation dizziness, discomfort, diarrhea, dysphagia, expectoration, emesis, little sputum, headache, expectoration, runny nose, malaise sputum, shortness of breath, muscular soreness, conjunctivitis, myalgias, sore throat, chills, chest, distress, fever, gasp, cough, dyspnea, cardiogenic shock, acute coronary syndrome, hypoxia, cardiopulmonary arrest, acute coronary syndrome, congestive heart failure, severe acute respiratory infection, acute kidney injury, septic shock, acute respiratory distress syndrome, acute myocardial infarction, pneumonia, hypoxia, anorexia, fatigue arrhythmia, asymptomatic, afebrile, cardiac arrhythmia, lives\_in\_Wuhan, travel\_history\_location, additional\_information, chronic\_disease\_binary, chronic\_disease, **Alive**

**三、實驗方法**

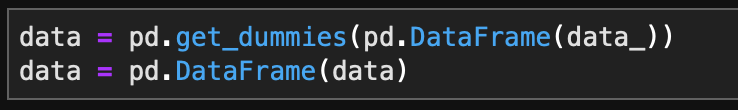
1. **NN**

從csv檔中讀入資料後，首先處理資料。

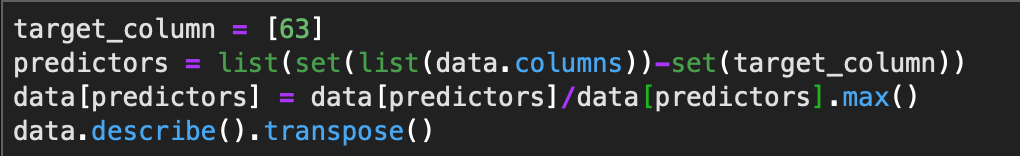
由於已經有離散化的age-range欄位，所以不使用原始的age欄位。

也不採用additional\_information欄位，因為該欄位為備註性質，資料內容比較瑣碎且雜亂，所以不採納為attribute。

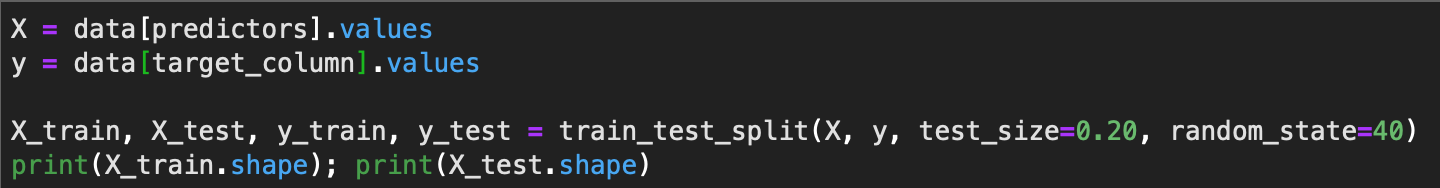
由於欄位值都不是numeric data，所以針對每個欄位做one hot encoding，使用pandas套件中的get\_dummies函數，產出一個做完one hot encoding的data frame。



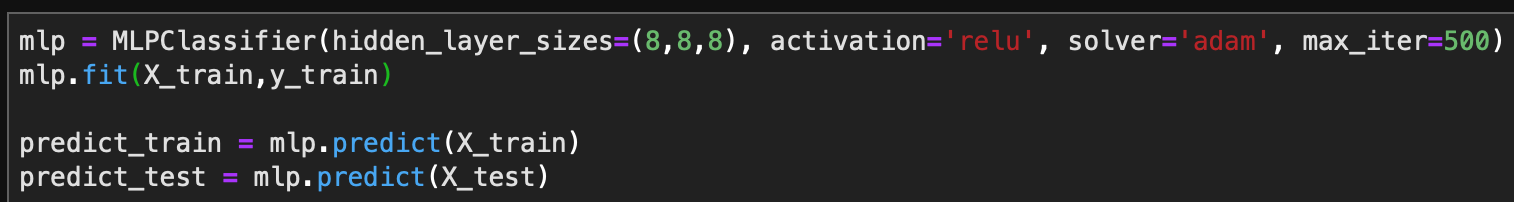
再來選定alive欄位為要預測的label，alive欄位中只有兩種結果，分別是1與0，分別代表病患生或死的狀態。在透過get\_dummies函數作用之後，新的data frame有64欄，alive欄位位於最後一欄，欄位名稱變為”63”，我們選定該欄位作為target\_column。並且將剩餘的欄位當作predictors，表示以這些欄位當作預測target\_column的養分，並且針對每一欄做標準化。



再來分割出training set and testing set，採用training set 80%，testing set 20%的比例。

****

接下來建構模型，我們使用的hidden layer架構為三層，每層分別有8個神經元，激勵函數使用的是relu，solver使用adam來優化權重，max iteration設為500次。

****

1. **SVM**

讀取csv檔後，因為SVM必須使用數值化的input&outpu，將欄位進一步做one hot encoding作為input，並將結果是否死亡(yes/no)轉為1/0。



接著將data切割為訓練集與測試集，使用的是test\_size = 0.3。

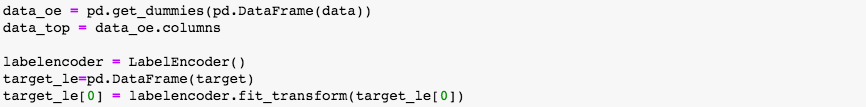


使用不同kernel(linear, rbf以及poly)和不同參數測試資料集，結果和分析如下方所示。



1. **Decision Tree & Random Forest**

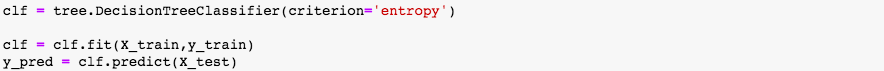
將 Data 經過 panda.get\_dummies 進行 One-hot-encoding 後產生 dataframe 作為訓練資料，target 則是用 LableEncoding 將結果 Yes 跟 No 標記為 1 和 0。



接著將資料分割出 80% training set 及 20% testing set 進行訓練。



這邊實作 Decision Tree 的分類器，criterion 選擇 entropy 的方式，random\_state 設為 1。



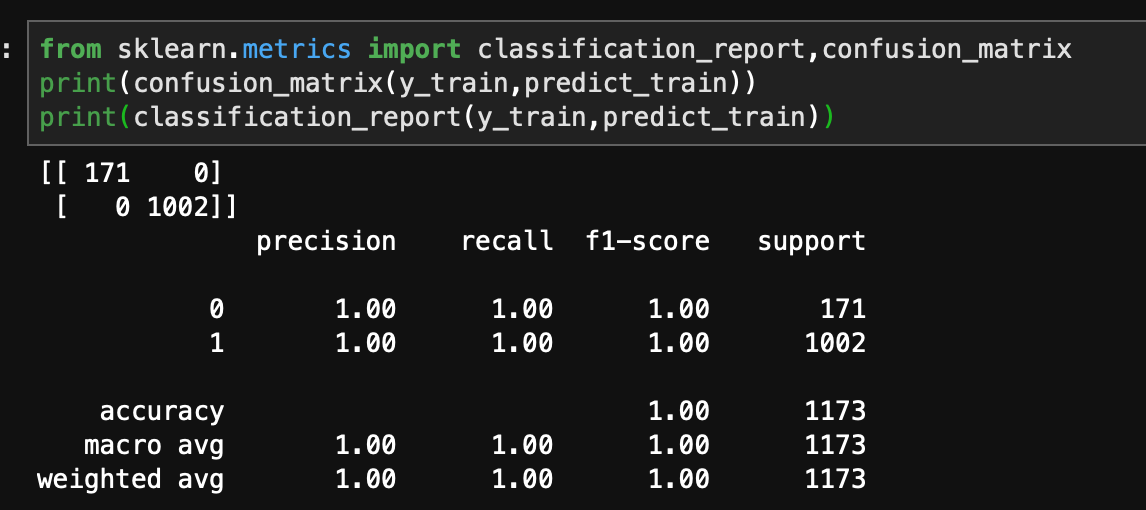
另外也實作 Random Forest 的分類器，criterion 一樣選擇 entropy 的方式 estimators 設為 1。



**四、實驗結果與分析**

1. **NN**

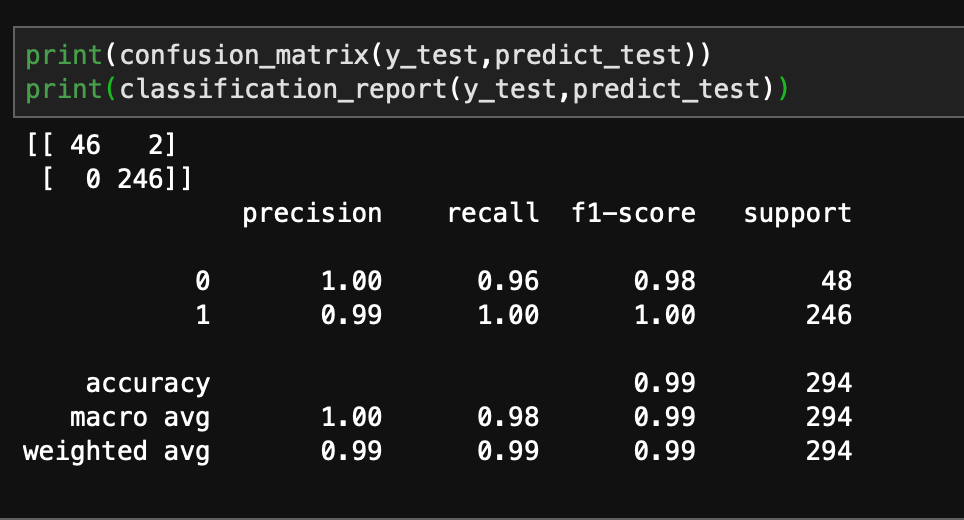
我們在training set上面得到的結果如下



Macro average precision為100%，Macro average F1-score為100%

Weighted average precision為100%，Weighted average F1-score為100%

而在testing set的結果如下



Macro average precision為100%，Macro average F1-score為99%

Weighted average precision為99%，Weighted average F1-score為99%

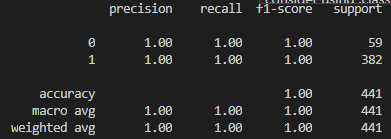
可以看出模型在training set跟testing set上面都表現得非常良好，我們認為這可能是因為dataset之中有明顯優勢的data，例如只要有武漢旅遊史、來自中國城市，或是徵狀裡面有發燒、咳嗽等狀況，相對其他屬性組合，會有更高的機率被判為死亡的狀態。

未來期望可以加入更多資料量，用於針對symtom做更進階的分析。

不過symtom相關的完整數據非常稀少，也是本次實作前置作業的困難之處。

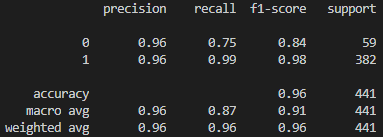
1. **SVM**

**kernel = linear**

****

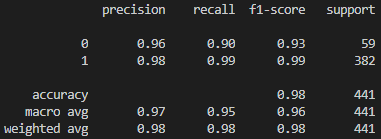
可以看見linear的kernel能100%準確地將測試資料集分類，顯示該資料集當中可能有dominant的欄位存在。而以現實所知的事實而言，確實是否住在武漢或是有沒有慢性病，都有極大可能降低存活率，因此linear kernel可以輕易劃分兩個類別，甚至在測試集當中辨識率能夠達到100%，表示此資料集真的非常集中。

**kernel = poly, degree = 5~12**

****

使用poly kernel調整degree = 1~20當中，正確率最高的結果落在5~12之間。

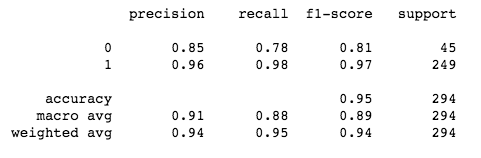
**kernel = rbf, gamma = 0.3~0.5**

****

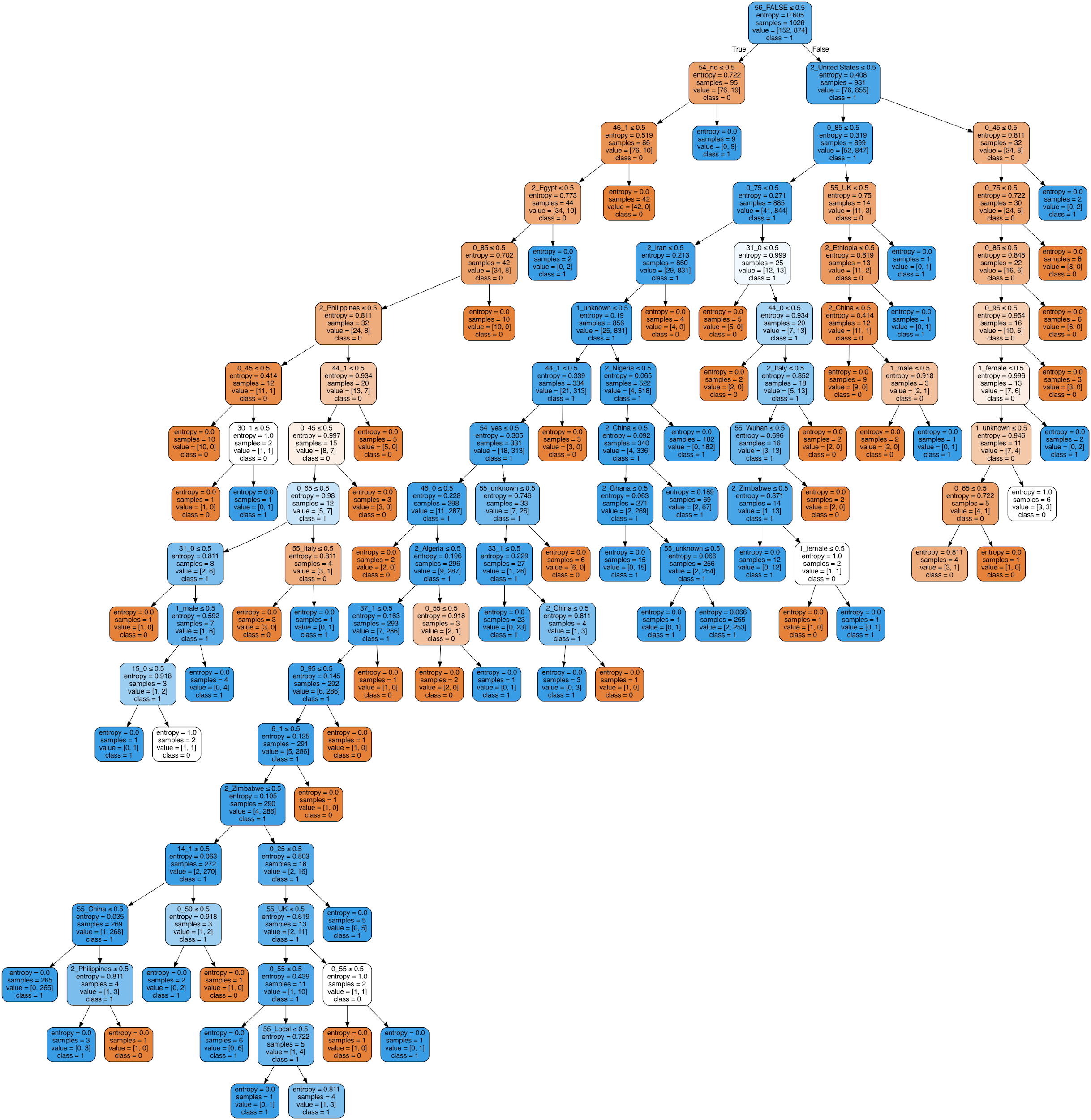
使用rbf調整參數時，gamma值越大代表越不影響附近資料點，最佳參數落在0.3~0.5之間可能代表資料集的同質性高、兩群異質性也很高，才能在較大影響範圍中得到98%正確率。

1. **Decision Tree & Random Forest**

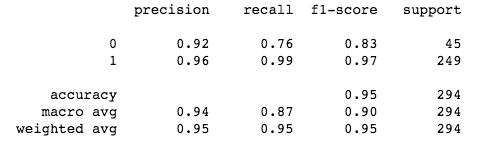
使用 metrics.classification\_report 來檢視訓練結果如下。Decision Tree 的部分，macro precision 為 91%，Macro average F1-score為 89%，Weighted average precision為 95 %，Weighted average F1-score為 95%。模型上的表現雖不如前面的 NN，但準確率也有達到 9 成算是不錯的結果。



將 Decision Tree 的樹狀圖透過 graphviz 繪製出來如下圖，第一個用來區分的Feature 為是否有慢性病，可以得知有無慢性病是武漢肺炎是否會造成死亡很重要的因素，而是否去過武漢以及居住地也是影響預測結果很重要的因素。



Random Forest 的部分一樣使用 metrics.classification\_report 來檢視訓練結果。macro precision 為 94%，Macro average F1-score為 90%，Weighted average precision為 95 %，Weighted average F1-score為 95%。



根據結果可以看到 Random Forest 的表現比 Decision Tree 略高一點點，但差異其實不大，我們推測可能是因為這個資料集的資料量不算龐大，且經過整理後資料的複雜度亦不高，只使用一個 Decision Tree 的結果已經足以訓練出不錯的結果了，因此在資料較不複雜的狀況下，其實可以採用 Decision Tree 來做預測即可。

1. **結論**

此次實驗資料貼近時事，在尋找資料集的過程也因為切身相關而更加想要進一步探索。原本我們想要尋找的是檢驗資料，想要透過自己的分析觀察，了解目前的檢驗狀況以及怎麼樣的個案可能會是陽性/陰性。但是卻發現要找到完整的資料集相當不容易。也許是保護當事人，各國檢驗資料幾乎不公開，找到的都是已確診的資料，來自各國的欄位及記錄方式也不盡相同。相較於分析結果，我們花了不少時間討論如何將有限資料作適當前處理。

而結果也還算驗證目前大眾資訊，就是慢性病和地區是病情是否致死相當重要的主因，也是一次有趣的結果。