實作聯邦式學習 -- 預測肺癌患者罹患二次癌症風險

1. 資料統整

1.1 資料來源

由衛生福利部統計台灣醫院肺癌患者的結果作為本次研究的資料集。其中包含長庚、中山、中慈、亞東、仁愛等五家醫院。年份為 2011 ~ 2020 共 10545 筆資料。每一家醫院各分別為一個獨立機構。

1.2 資料特徵

表一 特徵名稱與簡介

特徵名稱	簡介
肺癌類別（TMRPRMYST）	國際癌症病理和分類學的第三版
側性（SYMOGN）	
性別（Gender）	
肺癌的其一種類別（PTHLTYPE）	
SSF1	同側肺部的獨立腫瘤結節
SSF2	內臟胸膜侵犯/彈性層
SSF3	治療前的身體狀態評估
SSF4	惡性胸膜積液
SSF6	EGFR（表皮生長因子受體）基因突變
SSF7	ALK（Anaplastic Lymphoma Kinase）基因突變
確診時年齡	
分級/分化	病理學分化
腫瘤大小	
區域淋巴結侵犯數目	陽性區域淋巴結
整併期別	
原發部位手術邊緣	
放射性治療	
全身性治療	
BMI	
吸菸行為	
嚼檳榔行為	
喝酒行為	
手術切除原發部位	

2. 訓練過程

2.1 建立連線

利用 python 套件 flower 實作聯邦式學習預測肺癌患者罹患二次癌症的風險。並用 Ngrok 解決沒有固定 IP 的問題，處理完 gRPC 協議的問題後[1]，成功讓擁有不同機構資料的電腦可以進行連線和訓練。

2.2 資料不平衡

利用 flower 套件中資料權重的設定，將罹患二次癌症患者的錯誤比重放大，盡可能讓模型不傾向於預測患者不會罹患二次癌症。

3. 模型效能評斷

利用多種參數評斷模型的好壞。因為資料太過於不平衡，對於全部預測為不發生二次癌症的模型，正確率高達 95%。所以，除了 accuracy 之外，我們增加 precision、recall、f-score、mcc、AUC 等多種方式檢測。不僅如此，我們也更著重於 accuracy 不高，但 recall 值高的模型，希望能盡可能的預測有高度二次癌症風險的病患。

4. 特徵重要性解釋

利用特徵解釋工具 shap 解釋每個特徵的重要性，並以直方圖展示每個特徵對於模型輸出的貢獻程度。

5. 本次大專生計畫發想原由

在進行實際的專題研究過程中，我們注意到跨不同醫療機構的肺癌病患存在相同特徵卻有不同種類的情況，這引發了我們對病患特性背後潛在差異的好奇心。具體而言，我們選擇以仁愛醫院和長庚醫院為對象進行比較。

值得注意的是，在進行數據分析時我們發現，仁愛醫院的肺癌病患人數明顯較長庚醫院為少，且更為顯著的是，在仁愛醫院的肺癌病患中，並沒有出現 C340 類別的病患。這樣的現象引起了我們極大的興趣，並成為整個大專生計畫的發想原由。

6. 參考資料:

[1] Ghosh Bratin et, al. “Federated Learning Platform for Covid-19 Detection” [Online] Available: https://www.cs.hku.hk/images/Content/fyp-competition/FinalReport_FYP20060_BratinGhosh_3035437692.pdf

7. 專題製作

由我和張育安同學於 113 第一學期共同研究完成