資料探勘　專案 2
COVID-19 重症高風險族群分類分析

洪裕翔
國立成功大學　資訊工程研究所
碩士一年級
P76124215

1　資料

1.1　問題定義

自西元 2019 年底爆發以來，COVID-19 已成為全球嚴重的公共衛生危機。根據世界衛生組織的統計 [1]，截至 2023 年 11 月 12 日，全球 COVID-19 確診病例數已超過 7.7 億，死亡病例數則超過 690 萬。疾病的嚴重程度因人而異，有些人可能僅出現輕微症狀，但也有些人可能發展成重症，甚至導致死亡。

本專案以「識別 COVID-19 重症高風險族群」為主軸，關注 COVID-19 重症高風險族群的共同特徵，包括年齡較大、患有慢性病、免疫力低下等。這些族群感染 COVID-19 後，更容易出現重症病狀，可能需要住院治療或使用呼吸器。因此，若能在早期病程識別這些重症高風險族群，即可更早地積極醫療干預，降低重症發生率和死亡率。同時，這也有助於合理分配醫療資源，確保重症患者得到有效治療。

值得注意的是，除了 COVID-19，本專案也可應用在其他疾病上。這意味著本專案之研究成果不僅僅對當前的公共衛生挑戰有價值，更能擴展至未來可能出現的其他傳染性疾病。

1.2　資料設計

本節將定義用於生成資料集的特徵與規則，特徵共計 10 種，規則共計 5 種。本生成之資料集的預測目標為「此病人是否為 COVID-19 重症高風險族群」。

1.2.1　特徵

離散型

- 慢性病數量（$CHRONIC\_NUMBER$）
  - 範圍為 [0, 7]。
  - 整數型資料。
  - 參考之慢性病種類：

        * 癌症
        * 糖尿病
        * 心血管疾病
        * 肺病
        * 腎病
        * 肝病
        * 精神疾病
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.5, 0.2, 0.1, 0.08, 0.06, 0.03, 0.02, 0.01]。

- 性別（$GENDER$）
    - 男性為 1。
    - 女性為 0。
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.5, 0.5]。

- 懷孕（$IS\_PREGNANCY$）
    - 已懷孕或產後 6 週內為 1。
    - 未懷孕或非產後 6 週內則為 0。
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.9, 0.1]。

- 吸菸（IS_SMOKE）
    - 有吸菸或曾有吸菸為 1。
    - 無任何吸菸經驗則為 0。
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.7, 0.3]。

- 正常作息數量（$REGULAR\_LIFE\_NUMBER$）
    - 範圍為 [0, 5]。
    - 整數型資料。
    - 參考之正常作息種類：
        * 規律作息
        * 均衡飲食
        * 多喝水
        * 多運動

        * 適量補充營養素
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.4, 0.2, 0.2, 0.1, 0.05, 0.05]。
- 重複感染次數（$REINFECTION\_NUMBER$）
    - 範圍為 [0, 4]。
    - 整數型資料。
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.8, 0.1, 0.05, 0.03, 0.02]。
- 疫苗劑數（$VACCINE\_NUMBER$）
    - 範圍為 [0, 5]。
    - 整數型資料。
    - 抽樣方法為 Multinomial Distribution。
        * 機率定為 [0.1, 0.05, 0.05, 0.05, 0.5, 0.25]。

連續型

- 年齡（AGE）
    - 範圍為 [0 - 100]。
    - 小數型資料。
    - 抽樣分布定為 Normal Distribution。
        * 平均定為 43 歲 [2]。
        * 標準差定為 5.0。
- 身體質量指數（BMI）
    - 範圍為 [15 - 40]。
    - 小數型資料。
    - 抽樣分布定為 Normal Distribution。
        * 平均定為 21.0。
        * 標準差定為 0.5。
- 經濟（$ECONOMICS$）
    - 範圍為 [1 - 5]。
    - 小數型資料。
    - 抽樣分布定為 Normal Distribution。
        * 平均定為 3.0。
        * 標準差定為 1.0。

1.2.2 正確規則

以下皆為針對本專案的實作假設定義，並非真實情況。

COVID-19 重症高風險族群（$IS\_HIGH\_RICK\_FOR\_SEVERE\_COVID\_19 = 1$）

- 規則 1 - 滿足以下所有特徵（且，AND）：
  - $AGE \geq 50$
  - $VACCINE\_NUMBER \leq 2$
  - 滿足以下至少 3 項特徵（或，OR）：
    - $BMI \geq 30$
    - $CHRONIC\_NUMBER \geq 1$
    - $IS\_PREGNANCY = 1$
    - $IS\_SMOKE = 1$
    - $REGULAR\_LIFE\_NUMBER \leq 3$
- 規則 2 - 滿足以下所有特徵（且，AND）：
  - $10 < AGE < 50$
  - $VACCINE\_NUMBER \leq 2$
  - 滿足以下至少 3 項特徵（或，OR）：
    - $BMI \geq 30$
    - $CHRONIC\_NUMBER \geq 2$
    - $IS\_PREGNANCY = 1$
    - $IS\_SMOKE = 1$
    - $REGULAR\_LIFE\_NUMBER \leq 3$
    - $ECONOMICS \leq 2$
- 規則 3 - 滿足以下所有特徵（且，AND）：
  - $AGE \leq 10$
  - $VACCINE\_NUMBER \leq 2$
  - 滿足以下至少 2 項特徵（或，OR）：
    - $BMI \geq 30$
    - $CHRONIC\_NUMBER \geq 1$
    - $REGULAR\_LIFE\_NUMBER \leq 3$
    - $ECONOMICS \leq 2$
- 規則 4 - 滿足以下所有特徵（且，AND）：
  - $VACCINE\_NUMBER = 0$

– $REINFECTION\_NUMBER \geq 1$

非 COVID-19 重症高風險族群（$IS\_HIGH\_RICK\_FOR\_SEVERE\_COVID\_19 = 0$）

- 規則 5 - 不滿足所有 COVID 重症高風險族群規則。

1.2.3 其他規則

本生成之資料集有對以下特徵制定細節規則，以避免生成不符合現實狀況的資料。

- 懷孕（$IS\_PREGNANCY$）
  - 若性別（$GENDER$）為 0，且年齡（AGE）大於等於 15、小於等於 55，則以 1.2.1 節定義之抽樣方法決定此特徵之值。
  - 若不符合上點之條件，此特徵定義為 0。

1.3 生成結果

基於 1.2 節的定義，本專案生成共計 10000 筆資料之無干擾資料集，如表 1 所示。其中，有 9512 筆資料為非 COVID-19 重症高風險族群，有 488 筆資料為 COVID-19 重症高風險族群。

本專案亦有生成共計 10000 筆資料之干擾資料集，生成之規則定義為：將 1.2 節絕對正確規則之 1、2、3 的「$VACCINE\_NUMBER \leq 2$」特徵移除，其餘保持不變。如表 2 所示。其中，有 8238 筆資料為非 COVID-19 重症高風險族群，有 1762 筆資料為 COVID-19 重症高風險族群。

索引	慢性病數量	性別	懷孕	吸菸	正常作息數量	重複感染次數	疫苗劑數	年齡	身體質量指數	經濟
1	3	1	0	0	3	0	2	55.76	21.42	2.75
2	1	1	0	0	0	0	1	47.25	20.55	1.69
3	1	0	0	0	0	0	4	44.13	21.78	2.95
…										
9998	0	0	1	1	3	1	5	46.26	21.04	2.46
9999	1	0	1	1	2	0	0	41.89	21.11	4.14
10000	0	0	1	1	1	2	0	45.10	21.30	3.28

表 1: 生成之無干擾資料集範例。

索引	慢性病數量	性別	懷孕	吸菸	正常作息數量	重複感染次數	疫苗劑數	年齡	身體質量指數	經濟
1	3	1	0	1	0	0	4	42.00	20.71	3.24
2	1	0	0	0	1	0	4	50.01	20.99	2.19
3	0	1	0	0	0	0	4	40.85	22.03	3.89
…										
9998	5	0	0	0	2	0	4	44.94	21.49	3.04
9999	0	1	0	1	0	0	1	42.12	21.00	3.12
10000	0	0	0	0	1	0	4	48.52	21.71	2.98

表 2: 生成之干擾資料集範例。

2 分類模型

本專案中，除了指定的決策樹（Decision Tree）之外，我另外選擇了隨機森林（Random Forest）作為比較模型。各模型之執行結果請見 3.1 節。

3 分析

3.1 結果

3.1.1 決策樹（Decision Tree）

無干擾資料

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	1902
1	0.98	0.94	0.96	98
Accuracy			1.00	2000
Macro avg	0.99	0.97	0.98	2000
Weighted avg	1.00	1.00	1.00	2000

表 3: 決策樹於無干擾資料集之效能。

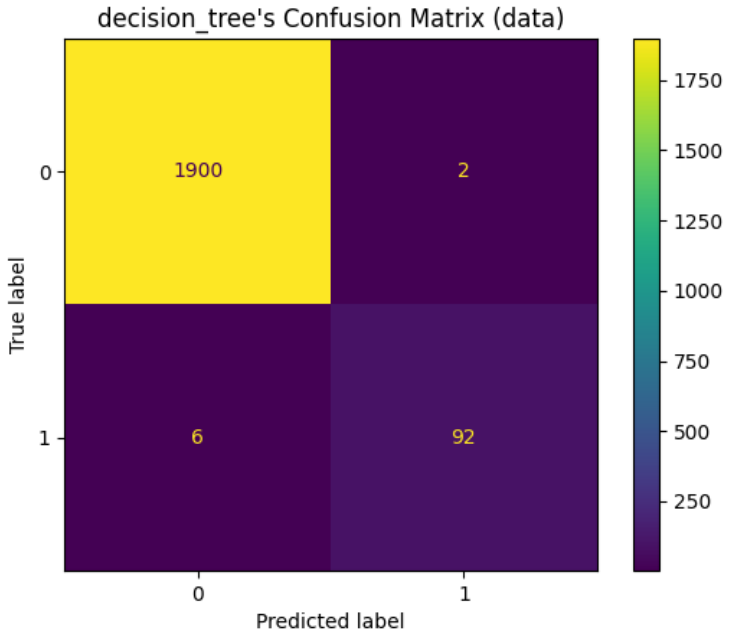

圖 1: 決策樹於無干擾資料集之混淆矩陣。

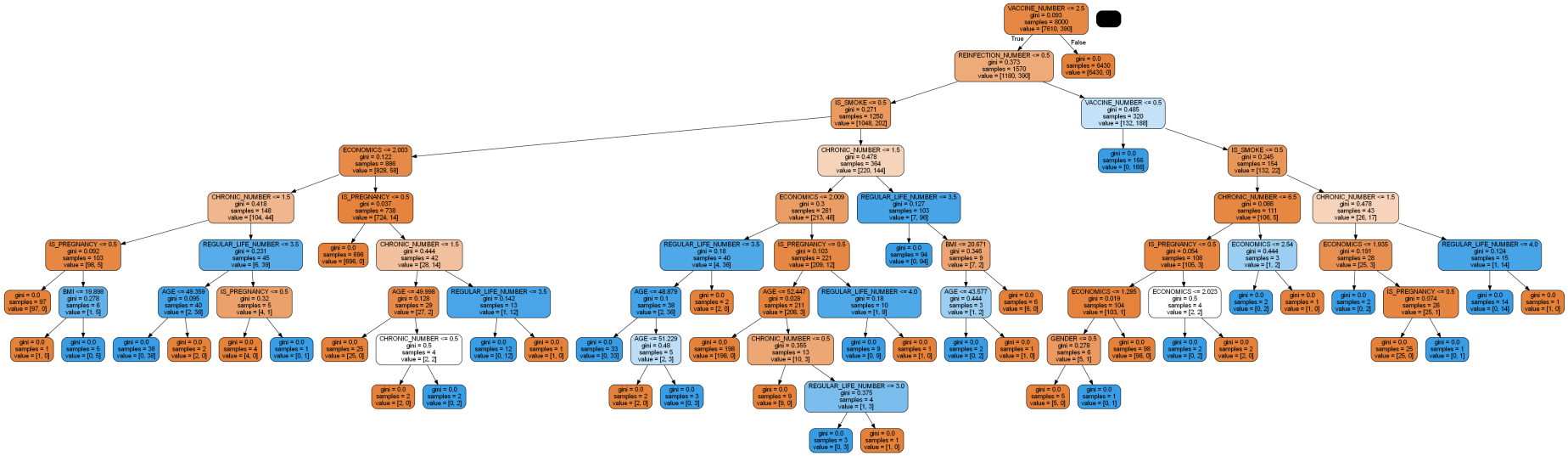

圖 2: 決策樹於無干擾資料集之結構。

干擾資料

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	1649
1	1.00	0.99	0.99	351
Accuracy			1.00	2000
Macro avg	1.00	0.99	0.99	2000
Weighted avg	1.00	1.00	1.00	2000

表 4: 決策樹於干擾資料集之效能。

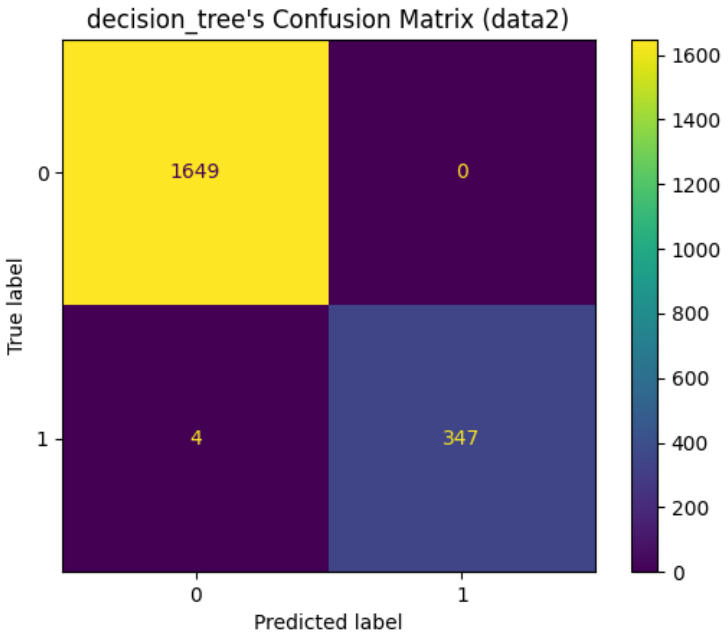

圖 3: 決策樹於干擾資料測試集之混淆矩陣。

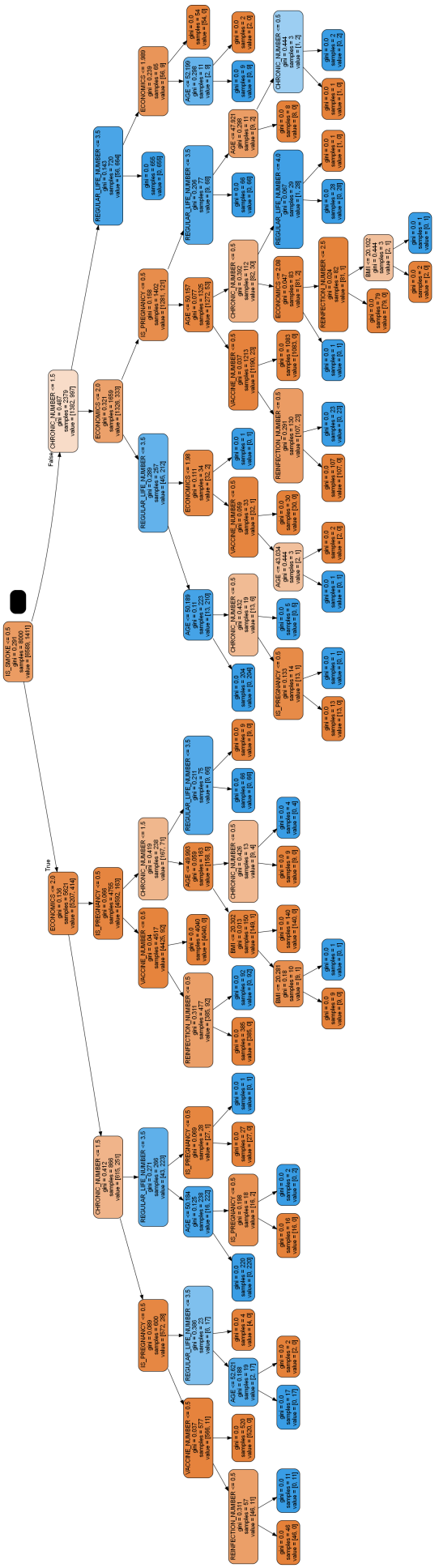

圖 4: 決策樹於干擾資料測試集之結構。

3.1.2 隨機森林（Random Forest）

無干擾資料

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	1905
1	0.97	0.91	0.93	95
Accuracy			0.99	2000
Macro avg	0.98	0.95	0.97	2000
Weighted avg	0.99	0.99	0.99	2000

表 5: 隨機森林於無干擾資料測試集之效能

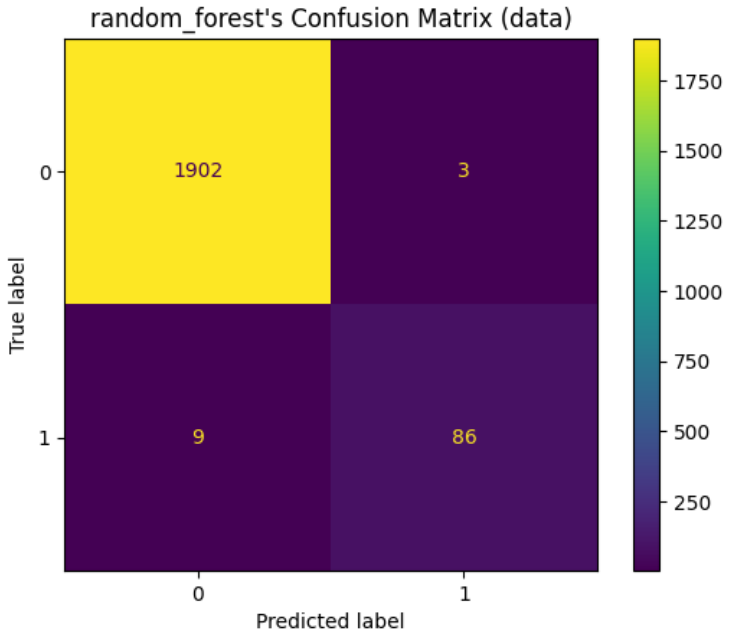

圖 5: 隨機森林於無干擾資料測試集之混淆矩陣。

干擾資料

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	1666
1	0.99	0.99	0.99	334
Accuracy			1.00	2000
Macro avg	1.00	0.99	0.99	2000
Weighted avg	1.00	1.00	1.00	2000

表 6: 隨機森林於干擾資料測試集之效能

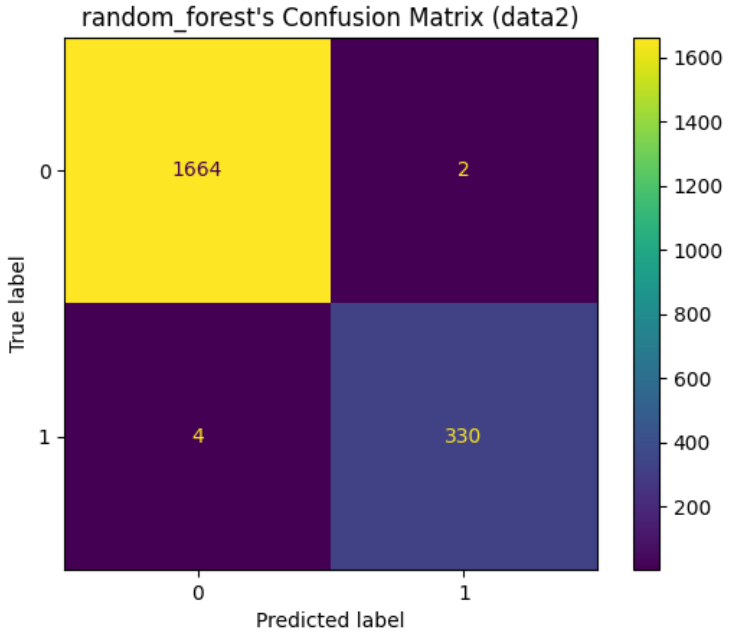

圖 6: 隨機森林於干擾資料測試集之混淆矩陣。

3.2 比較

本節主要比較 1.2.2 節定義之正確規則（下稱定義規則）與決策樹根據未干擾資料集、干擾資料集生成之規則（下稱未干擾分類規則、干擾分類規則）的差異。

首先探討無干擾資料集之結果，由圖 2 可以發現，決策樹最上層、次層節點的判斷條件都是定義規則的重要特徵（疫苗劑數、重複感染次數），但之後的下層節點卻沒有使用剩下的重要特徵（年齡）繼續分裂，反而以需同時滿足數個才成立的次級特徵（抽菸、經濟、慢性病數量等等）建樹，年齡集中於決策樹的下層、接近末端的位置才被用於判斷。

基於上述結果，在干擾資料集的設計上，我移除了定義規則中，規則 1、2、3 的「疫苗劑數」這項重要特徵，以觀察決策樹節點條件之變化。由圖 4 能觀察到，決策樹最上層、次層節點的判斷條件皆不再使用疫苗劑數、重複感染次數等重要特徵。然而，決策樹並沒有選擇年齡替代疫苗劑數，反而仍選擇次級特徵（抽菸、經濟、慢性病）作為上層節點的判斷條件，且可以發現圖 4 次級特徵的結構和條件和圖 2 非常相似，中下層的節點才有較大的差異，而年齡仍然位於決策樹的中下層節點。

上面的發現表明，決策樹在面對缺失某些重要特徵的情況下，會傾向於保留相似的次級特徵結構，而非直接替換為其他類似重要性的特徵。因此，即便在不同資料集條件下，決策樹的生成結構在最上層和次層節點仍然保持一致，差異主要體現在中下層節點的條件上。

3.3 討論

本節主要討論未干擾資料集與干擾資料集在各模型上的結果差異和可能原因。

由表 3、4、5、6 可以發現 2 個現象：

- 決策樹的效能表現在 2 種資料集上都較隨機森林好
- 干擾資料的預測結果在 2 種模型上都較無干擾資料好

3.3.1 決策樹的效能表現在 2 種資料集上都較隨機森林好

這個現象和理論上的認知不太相符，但對於此現象，我推測可能由以下原因造成：

- 訓練過程沒有發生過擬合
  相對於決策樹，隨機森林的其中一項優勢是：面對可能發生過擬合的情況時，其能通過集成多個決策樹的結果，避免發生過擬合。換句話說，若現在的訓練任務不會發生過擬合，隨機森林便無法充分發揮其設計理念之優勢，在這樣的情況下，決策樹是有機會表現得比隨機森林好的。

- 資料與模型特性
  比較決策樹和隨機森林的設計理念與原理，決策樹基於給定的全部資料建樹，隨機森林則先通過隨機抽樣生成多種小資料集，再用以訓練多個弱決策樹。有可能本專案的資料集特性不適合被隨機抽樣，也就是包含全資料的決策樹比起隨機森林，更能適應資料特性，才造成這樣的現象。

- 超參數調整
  本專案使用的決策樹與隨機森林之超參數皆為預設值，預設的超參數不一定是此任務的最佳值，因此存在調整隨機森林超參數後，使其效能反超決策樹的可能性。

3.3.2 干擾資料的預測結果在 2 種模型上都較無干擾資料好

干擾資料與無干擾資料的唯一差異在於：干擾資料移除了在 COVID-19 重症高風險族群的定義 1、2、3 中的「$VACCINE\_NUMBER \leq 2$」特徵。對此，我認為增加或刪減特徵對於模型的影響是複雜的，尤其本專案的 2 個模型都基於決策樹。決策樹的分裂基於特徵的訊息增益、基尼不確定性等標準，增加特徵會提供更多的分裂選擇，但更多的分裂選擇不一定意味著對訓練、預測有益。

除此之外，資料平衡性也可能是影響預測結果的原因之一，由 1.3 節可知，移除部份特徵後得到的干擾資料，符合 COVID-19 重症高風險族群的資料數量大幅提高，也就是整體資料的平衡性提昇了，這亦有可能是造成此節討論之現象的原因之一。

References

[1] Who coronavirus (covid-19) dashboard — who coronavirus (covid-19) dashboard with vaccination data. `https://covid19.who.int`. Accessed: 2023-11-12.

[2] 人口金字塔- 人口推估統計查詢系統- 國家發展委員會. `https://pop-proj.ndc.gov.tw/pyramid.aspx?uid=64&pid=60`. Accessed: 2023-11-12.