新冠肺炎初步分析

韓明澄

December 22, 2024

1 COVID-19 資料集展示

下圖展示了 COVID-19 患者資料的一部分特徵。資料集包括患者的基本資訊、健康狀況、診療機構等欄位，來自於 Kaggle 的公開資料集[1]。

USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	DATE_DIED	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU
2	1	1	1	03/05/2020	97	1	65	2	2	2	2	2	1	2	2	2	2	2	3	97
2	1	2	1	03/06/2020	97	1	72	97	2	2	2	2	1	2	2	1	1	2	5	97
2	1	2	2	09/06/2020	1	2	55	97	1	2	2	2	2	2	2	2	2	2	3	2
2	1	1	1	12/06/2020	97	2	53	2	2	2	2	2	2	2	2	2	2	2	7	97
2	1	2	1	21/06/2020	97	2	68	97	1	2	2	2	1	2	2	2	2	2	3	97
2	1	1	2	9999-99-99	2	1	40	2	2	2	2	2	2	2	2	2	2	2	3	2
2	1	1	1	9999-99-99	97	2	64	2	2	2	2	2	2	2	2	2	2	2	3	97
2	1	1	1	9999-99-99	97	1	64	2	1	2	2	1	1	2	2	2	1	2	3	97
2	1	1	2	9999-99-99	2	2	37	2	1	2	2	2	1	2	2	1	2	2	3	2
2	1	1	2	9999-99-99	2	2	25	2	2	2	2	2	2	2	2	2	2	2	3	2
2	1	1	1	9999-99-99	97	2	38	2	2	2	2	2	2	2	2	2	2	2	3	97
2	1	2	2	9999-99-99	2	2	24	97	2	2	2	2	2	2	2	2	2	2	3	2

圖 1: COVID-19 資料集範例

1.1 COVID-19 患者資料初步介紹

這筆資料包含了 COVID-19 患者的多項特徵，涵蓋了患者的基本屬性、健康狀況、治療方式、診斷結果及死亡情況等。以下是資料的各項欄位的初步介紹：

- **基本資訊**
  - **sex**：患者性別。1 表示女性，2 表示男性。
  - **age**：患者年齡（以數字表示）。
  - **classification**：COVID-19 的檢測結果分類。1-3 表示不同程度的確診，4 或以上則表示患者未感染 COVID-19 或檢測結果不確定。

[1]資料來源：`https://www.kaggle.com/datasets/meirnizri/covid19-dataset`

  - **patient type**：患者在診療機構的治療類型。1 表示回家療養，2 表示住院。

- **健康狀況與病史**
  - **pneumonia**：患者是否已經患有肺炎。
  - **pregnancy**：患者是否為孕婦。
  - **diabetes**：患者是否患有糖尿病。
  - **copd**：患者是否有慢性阻塞性肺病（COPD）。
  - **asthma**：患者是否有哮喘病史。
  - **inmsupr**：患者是否為免疫抑制狀態。
  - **hypertension**：患者是否有高血壓。
  - **cardiovascular**：患者是否有心血管疾病。
  - **renal chronic**：患者是否患有慢性腎病。
  - **other disease**：患者是否患有其他疾病。
  - **obesity**：患者是否有肥胖問題。
  - **tobacco**：患者是否為煙草使用者（吸煙者）。
- **診療及醫療設施資訊**
  - **usmr**：患者是否在一級、二級或三級的醫療單位接受治療。
  - **medical unit**：患者接受治療的醫療單位類型（如國家健康系統中的不同機構）。
- **重症情況**
  - **intubed**：患者是否已連接呼吸器。
  - **icu**：患者是否已被送入重症加護病房（ICU）。
- **結局**
  - **date died**：患者死亡日期。若患者未死亡則為 `9999-99-99`。

2 資料初步分析方向

可以根據不同變數進行以下初步分析：

1. **描述性統計**：檢查年齡、性別、吸煙情況等特徵的分佈。

2. **健康狀況分析**：檢查健康狀況變數（如糖尿病、肥胖、慢性腎病等）與 COVID-19 症狀嚴重程度（如住院、是否使用呼吸器）的關聯。
3. **死亡分析**：以 `date died` 欄位來判斷死亡情況，分析哪些變數與死亡之間具有顯著性關聯。
4. **重症因素**：觀察 `intubed` 和 `icu` 這些重症相關變數與患者健康背景之間的關聯性，了解哪些群體更容易進入 ICU 或使用呼吸器。

這些分析有助於了解 COVID-19 患者的健康狀況、治療方式與死亡風險之間的關係。可以根據這些初步方向進一步深入數據分析，找出具體的關聯性和趨勢。

2.1 資料概述

分析包含多項影響 COVID-19 死亡風險的變數，包含人口統計（如年齡、性別）、健康狀況（如糖尿病、高血壓等）、生活習慣（如吸煙）等多個因素。我們基於描述性統計和各變數的死亡率柱狀圖進行初步分析。

描述性統計

變數	計數	平均值	標準差	最小值	25%	50%	75%	最大值
SEX	1048575	1.50	0.50	1	1	1	2	2
AGE	1048575	41.79	16.91	0	30	40	53	121
PATIENT_TYPE	1048575	1.19	0.39	1	1	1	1	2
PNEUMONIA	1048575	3.35	11.91	1	2	2	2	99
PREGNANT	1048575	49.77	47.51	1	2	97	97	98
DIABETES	1048575	2.19	5.42	1	2	2	2	98
COPD	1048575	2.26	5.13	1	2	2	2	98
ASTHMA	1048575	2.24	5.11	1	2	2	2	98
INMSUPR	1048575	2.30	5.46	1	2	2	2	98
HIPERTENSION	1048575	2.13	5.24	1	2	2	2	98
OTHER_DISEASE	1048575	2.44	6.65	1	2	2	2	98
CARDIOVASCULAR	1048575	2.26	5.19	1	2	2	2	98
OBESITY	1048575	2.13	5.18	1	2	2	2	98
RENAL_CHRONIC	1048575	2.26	5.13	1	2	2	2	98
TOBACCO	1048575	2.21	5.32	1	2	2	2	98
DIED	1048575	0.34	0.61	0	0	0	1	1

表 1: 描述性統計表

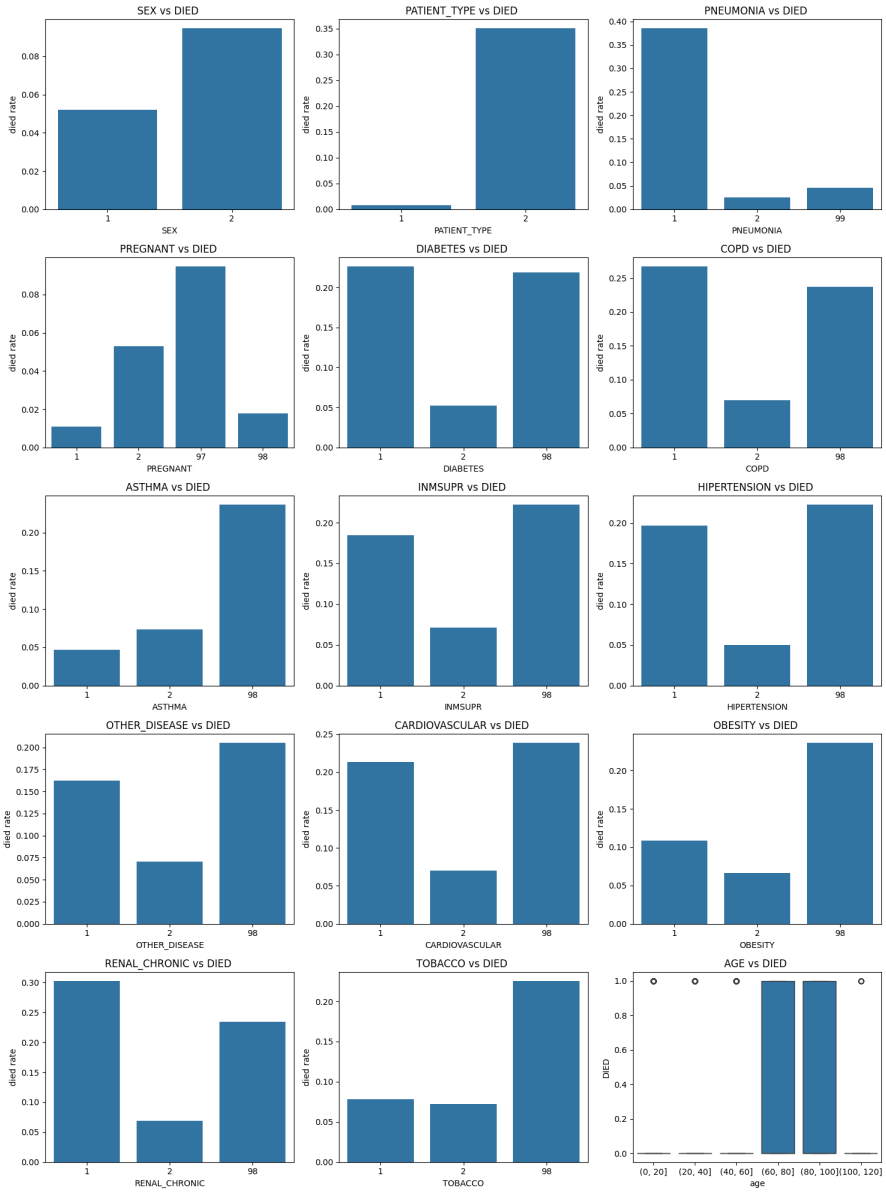

圖 2: 各變數對死亡率的影響

分析結果

根據圖 2 和表 1 的結果，對各變數的死亡率影響進行了如下分析：

- **SEX（性別）**：男性（2）的死亡率高於女性（1），可能表明男性在 COVID-19 中的死亡風險較高。
- **PATIENT_TYPE（患者類型）**：住院患者（2）的死亡率顯著高於回家療養的患者（1），這合理反映了住院患者的病情可能更嚴重。
- **PNEUMONIA（肺炎）**：患有肺炎的患者（1）的死亡率明顯高於無肺炎的患者（2），顯示出肺炎是 COVID-19 患者死亡風險的重要因素。
- **PREGNANT（懷孕狀態）**：孕婦（1）的死亡率低於非孕婦（2），但標記為「未知」（97 和 98）部分需要進一步清理和分析。
- **DIABETES（糖尿病）**：糖尿病患者（1）有較高的死亡率，這與糖尿病在 COVID-19 患者中的高風險一致。
- **COPD（慢性阻塞性肺病）**：患有 COPD 的患者（1）死亡率明顯較高，這是已知的 COVID-19 高風險因素。
- **ASTHMA（哮喘）**：哮喘患者（1）死亡率略高，但與無哮喘患者的差異不顯著。
- **INMSUPR（免疫抑制）**：免疫抑制患者（1）死亡率較高，顯示免疫抑制可能增加死亡風險。
- **HIPERTENSION（高血壓）**：高血壓患者（1）死亡率較高，說明高血壓與 COVID-19 患者的死亡風險相關。
- **OTHER_DISEASE（其他疾病）**：有其他疾病的患者（1）死亡率較高，說明合併其他疾病可能會增加死亡風險。
- **CARDIOVASCULAR（心血管疾病）**：心血管疾病患者（1）的死亡率較高，這與心血管疾病對 COVID-19 病情惡化的潛在影響一致。
- **OBESITY（肥胖）**：肥胖患者（1）的死亡率高於正常體重者，這是 COVID-19 中常見的風險因素。
- **RENAL_CHRONIC（慢性腎病）**：患有慢性腎病的患者（1）死亡率較高，這是 COVID-19 患者中的一個顯著風險因素。
- **TOBACCO（吸煙）**：吸煙者（1）和非吸煙者的死亡率差異不大，在此數據集中吸煙與 COVID-19 死亡風險沒有顯著關聯。
- **AGE（年齡）**：年齡顯著影響 COVID-19 死亡率，尤其在 60 歲以上，死亡率顯著上升。年齡是已知的 COVID-19 死亡風險因素。

根據以上分析結果，我們發現年齡、性別和多種健康狀況（如糖尿病、COPD、高血壓等）均對 COVID-19 患者的死亡率有顯著影響。這些結果可以幫助我們識別高風險群體，進而制定更具針對性的公共衛生策略。

2.2 相關性分析

圖 3 顯示了變數之間的相關性。該相關性熱圖提供了變數之間的直觀關聯性，其中較深的紅色表示更高的正相關，藍色表示負相關，白色表示接近零的相關性。以下是關鍵觀察點：

- **性別（SEX）與懷孕（PREGNANT）**：性別和懷孕狀態高度負相關（-0.99），因為懷孕僅適用於女性。
- **多種健康狀況的高度相關**：如糖尿病（DIABETES）、高血壓（HIPERTENSION）、心血管疾病（CARDIOVASCULAR）等相互正相關，這些共病狀況可能在 COVID-19 患者中同時出現。
- **死亡情況（DIED）與其他變數的相關性較低**：各變數與死亡情況的相關性都較低，顯示死亡風險可能受到多重因素的綜合影響。

邏輯回歸分析結果

表 2 展示了對死亡風險進行邏輯回歸分析的結果。模型中包含了多個解釋變數，每個變數的係數（coef）表明該變數對死亡風險的影響方向及程度，標準誤（std err）反映了估計的不確定性，z 值和 p 值則說明了係數的顯著性。95% 信賴區間（CI）顯示了係數的變異範圍，若此範圍不包含 0，則表明該變數對死亡風險有顯著影響。

數值解釋與發現

- **常數項（const）**：係數為 -11.7472，顯著性高（$p < 0.001$），表明在控制其他變數影響的情況下，基礎死亡風險較低。
- **性別（SEX）**：係數為 1.0327，顯示男性比女性的死亡風險更高。此係數意味著，若其他條件相同，男性的死亡風險是女性的約 2.81 倍（$e^{1.0327} \approx 2.81$），顯著性高（$p < 0.001$），95% 信賴區間為 [0.772, 1.294]。
- **年齡（AGE）**：係數為 0.0455，表明每增加一歲，死亡風險增長 4.6%（$e^{0.0455} \approx 1.046$），顯著性高（$p < 0.001$）。這表明年齡是 COVID-19 死亡風險的重要影響因素。

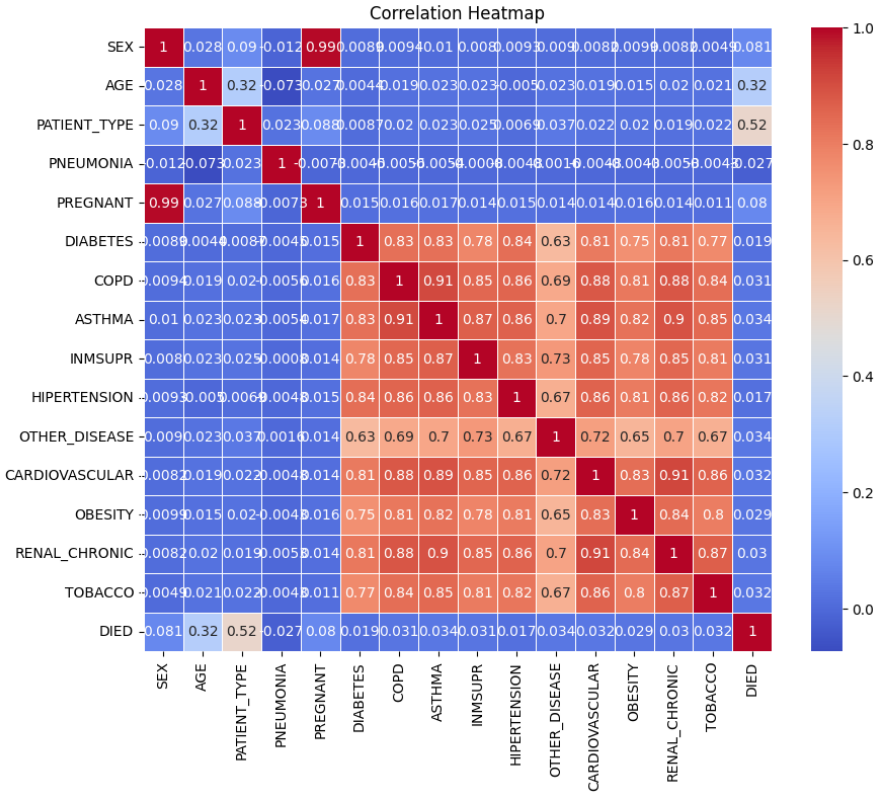

圖 3: COVID-19 變數的相關性熱圖

- **患者類型（PATIENT_TYPE）**：係數為 3.6678，顯示住院患者的死亡風險比非住院患者高很多（約 39 倍，$e^{3.6678} \approx 39$），顯著性極高（$p < 0.001$），顯示住院患者通常病情更嚴重，死亡風險相對更高。
- **糖尿病（DIABETES）**：係數為 -0.0048，但具有顯著性（$p < 0.001$），且 95% 信賴區間為 [-0.007, -0.002]。此結果顯示糖尿病患者的死亡風險在本模型中稍有減少，這可能是因為糖尿病患者通常與其他風險因素共同影響。
- **慢性阻塞性肺病（COPD）**：係數為 0.0072，顯著性高（$p < 0.001$），顯示 COPD 會稍微增加死亡風險。此結果表明呼吸系統疾病是增加 COVID-19 死亡風險的潛在因素。
- **哮喘（ASTHMA）**：係數為 0.0075，顯著性高（$p < 0.001$），意味著哮喘患者的死亡風險也有所增加，顯示慢性呼吸道問題對 COVID-19 死亡風險的影響。
- **肥胖（OBESITY）**：係數為 0.0028，具有顯著性（$p < 0.01$），表明肥胖對死亡風險有小幅增加的影響，但相對其他變數影響較小。

- **吸煙（TOBACCO）**：係數為 0.0051，顯著性高（$p < 0.001$），顯示吸煙者的死亡風險稍有增加，但影響相對有限。

表 2: COVID-19 死亡風險的邏輯回歸結果

變數	係數 (coef)	z 值	p 值	95% 信賴區間
常數項 (const)	-11.7472	-88.032	0.000	[-12.009, -11.486]
性別 (SEX)	1.0327	7.754	0.000	[0.772, 1.294]
年齡 (AGE)	0.0455	156.677	0.000	[0.045, 0.046]
患者類型 (PATIENT_TYPE)	3.6678	274.924	0.000	[3.642, 3.694]
肺炎 (PNEUMONIA)	-0.0120	-25.551	0.000	[-0.013, -0.011]
懷孕 (PREGNANT)	-0.0065	-4.640	0.000	[-0.009, -0.004]
糖尿病 (DIABETES)	-0.0048	-3.702	0.000	[-0.007, -0.002]
慢性阻塞性肺病 (COPD)	0.0072	4.401	0.000	[0.004, 0.010]
哮喘 (ASTHMA)	0.0075	4.061	0.000	[0.004, 0.011]
免疫抑制 (INMSUPR)	-0.0030	-2.115	0.034	[-0.006, -0.000]
肥胖 (OBESITY)	0.0028	2.708	0.007	[0.001, 0.005]
慢性腎病 (RENAL_CHRONIC)	-0.0025	-1.399	0.162	[-0.006, 0.001]
吸煙 (TOBACCO)	0.0051	3.779	0.000	[0.002, 0.008]

變數之間的散佈圖矩陣

圖 4 展示各變數之間的散佈圖矩陣。能夠直觀地顯示變數之間的分佈特徵和相互關係：

- **高共病率的健康狀況**：如糖尿病與高血壓之間存在一定的相關性。
- **性別（SEX）與懷孕（PREGNANT）之間的分佈**：性別與懷孕分佈顯示了邏輯上的明顯關聯。
- **死亡情況與其他變數的分佈特徵**：死亡與多個變數的分佈模式並不顯著，表明死亡風險可能受多重因素綜合影響。
- **慢性腎病** ：係數影響並不顯著，因此將其移除分析。

2.3 變數的共線性分析

根據 VIF（Variance Inflation Factor）表格，我們可以進一步分析這些變數之間的共線性問題。VIF 值表示每個變數與其他變數之間的多重共線性程度。一般而言，VIF 值大於 10 表示該變數可能存在顯著的多重共線性，而 VIF 值接近 1 表示該變數與其他變數之間的共線性較低。

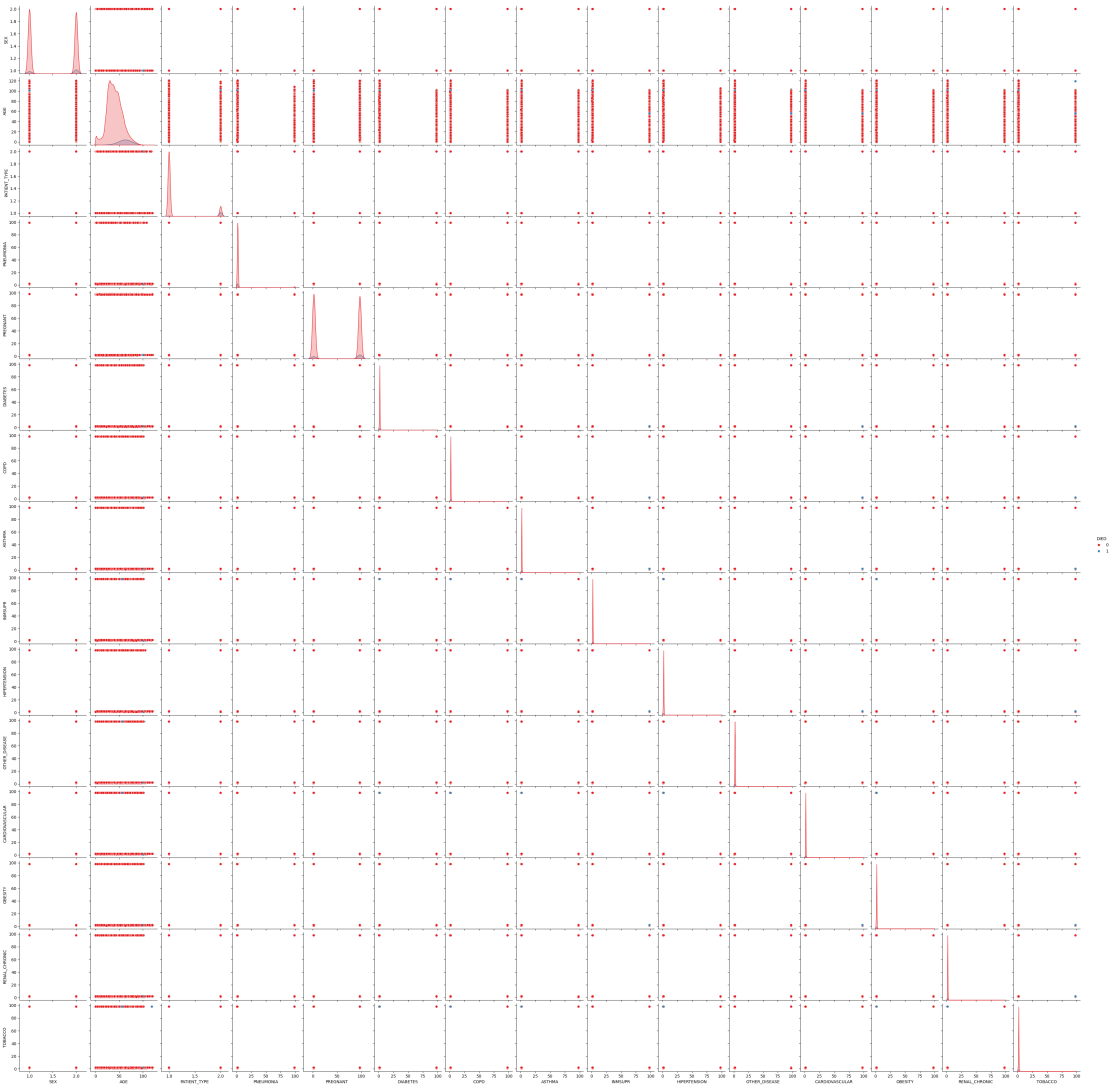

圖 4: COVID-19 各變數之間的散佈圖矩陣

表 3: VIF 共線性分析表

變數	VIF 值	解釋
常數項 (const)	278.49	極高共線性，可能由截距項影響
性別 (SEX)	69.21	高共線性，與懷孕狀態高度相關
年齡 (AGE)	1.12	無顯著共線性
患者類型 (PATIENT_TYPE)	1.12	無顯著共線性
肺炎 (PNEUMONIA)	1.01	無顯著共線性
懷孕 (PREGNANT)	69.20	高共線性，與性別高度相關
糖尿病 (DIABETES)	4.13	中度共線性
慢性阻塞性肺病 (COPD)	7.67	中度共線性
哮喘 (ASTHMA)	8.88	中度共線性
免疫抑制 (INMSUPR)	5.45	中度共線性
高血壓 (HIPERTENSION)	5.79	中度共線性
其他疾病 (OTHER_DISEASE)	2.34	較低共線性
心血管疾病 (CARDIOVASCULAR)	8.47	中度共線性
肥胖 (OBESITY)	4.13	中度共線性
腎臟慢性病 (RENAL_CHRONIC)	9.01	接近顯著共線性
吸煙 (TOBACCO)	4.97	中度共線性

分析結果

1. **高 VIF 值變數**：
    - **常數項（const）**：VIF 值高達 278.49，這是由於截距項的設置，使模型中所有變數的偏差疊加在常數項上。
    - **性別（SEX）和懷孕狀態（PREGNANT）**：VIF 值分別為 69.21 和 69.20，顯示這兩個變數具有非常高的共線性，這是因為懷孕僅適用於女性，導致性別和懷孕之間存在高度相關性。因此，在模型中納入這兩個變數會導致共線性問題。
2. **中度 VIF 值變數**：
    - **COPD（慢性阻塞性肺病）、哮喘（ASTHMA）、免疫抑制（INMSUPR）、高血壓（HIPERTENSION）、心血管疾病（CARDIOVASCULAR）、腎臟慢性病（RENAL_CHRONIC）**：這些變數的 VIF 值在 5 到 10 之間，表明這些變數之間可能存在一定的共線性。這些變數大多為慢性病，臨床上這些疾病通常會相互共現，因此在統計模型中表現出一定程度的多重共線性。
3. **低 VIF 值變數**：

- **年齡（AGE）、患者類型（PATIENT_TYPE）、肺炎（PNEUMONIA）**：VIF 值均接近 1，顯示這些變數之間的共線性不明顯。這意味著它們相對獨立，與其他變數的相關性較低，因此在模型中可以更直接地解釋其對目標變數（死亡率）的影響。

考慮解決方案

1. **移除或合併高共線性變數**：考慮移除性別或懷孕變數中的一個，以減少由這兩個高度共線的變數帶來的冗餘。由於懷孕僅適用於女性，可以考慮將性別作為控制變數，僅在特定分析中納入懷孕變數。
2. **考慮 PCA 方法處理具有臨床相關性的變數**：將慢性阻塞性肺病（COPD）、哮喘（ASTHMA）、高血壓（HIPERTENSION）等慢性病變數進行合併，以減少多重共線性，並增強模型的穩定性。例如，設置一個「合併健康狀況」變數，代表有無慢性疾病的風險，以簡化模型。
3. **使用正則化方法**：在模型中使用正則化（例如 Lasso 或 Ridge）方法，可以抑制多重共線性，並自動選擇具有最大解釋力的變數。Lasso 可以縮小部分變數的係數至零，從而自動進行變數選擇；而 Ridge 回歸則能有效減少多重共線性引起的估計不穩定性。
4. **進行主成分分析（PCA）**：對於高共線性的變數群，可以考慮使用主成分分析（PCA）來提取關鍵組合變數，並使用這些組合變數來替代原始變數。PCA 可以減少變數數量，並且保持原數據的主要變異，從而改善模型的穩定性。

3 主成分分析 (PCA) 結果

主成分分析是一種將高維數據降維的方法，用於提取數據的主要特徵，保持數據中的大部分變異性。在 COVID-19 數據中，我們應用了 PCA 來分析各特徵的重要性，並以視覺化形式展示其結果。

3.1 累積解釋變異比例

根據圖 5與表 4，可以看出第一主成分捕捉了 54.86% 的變異，前五個主成分總計解釋了 88.18% 的變異，顯示主成分分析有效降低了數據維度，同時保留了大部分的資訊量。

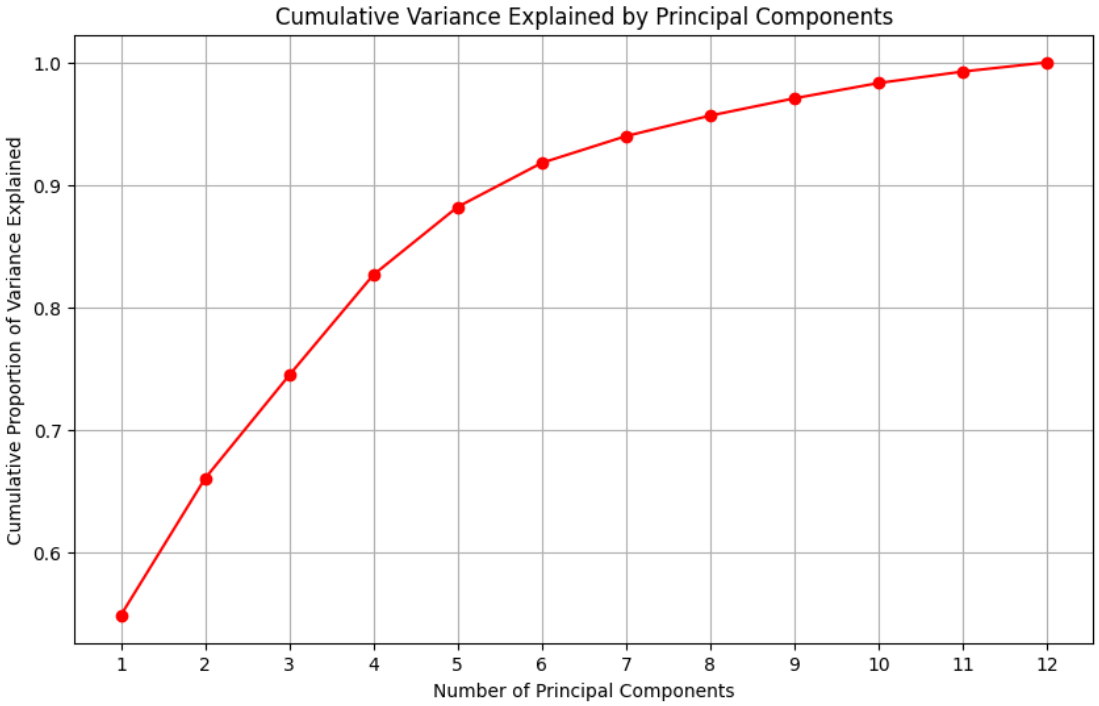

圖 5: PCA 主成份分析

表 4: 主成分解釋變異比例

主成分	解釋變異比例	累積解釋變異比例
第一主成分	54.86%	54.86%
第二主成分	11.17%	66.03%
第三主成分	8.43%	74.46%
第四主成分	8.20%	82.66%
第五主成分	5.54%	88.18%

3.2 PCA 負荷量分析

圖 6 為 PCA 的熱圖，展示各特徵對應主成分的影響力大小，表 5 則提供了詳細數值。第一主成分主要由慢性病特徵（如糖尿病、慢性阻塞性肺病等）驅動，而第二主成分與年齡和患者類型關聯密切。

3.3 PCA 點散圖

圖 7 為前兩個主成分的點散圖。藍色點代表未死亡患者 (DIED=0)，紅色點代表死亡患者 (DIED=1)。可以觀察到，雖然存在一定程度的重疊，但死亡患者的分

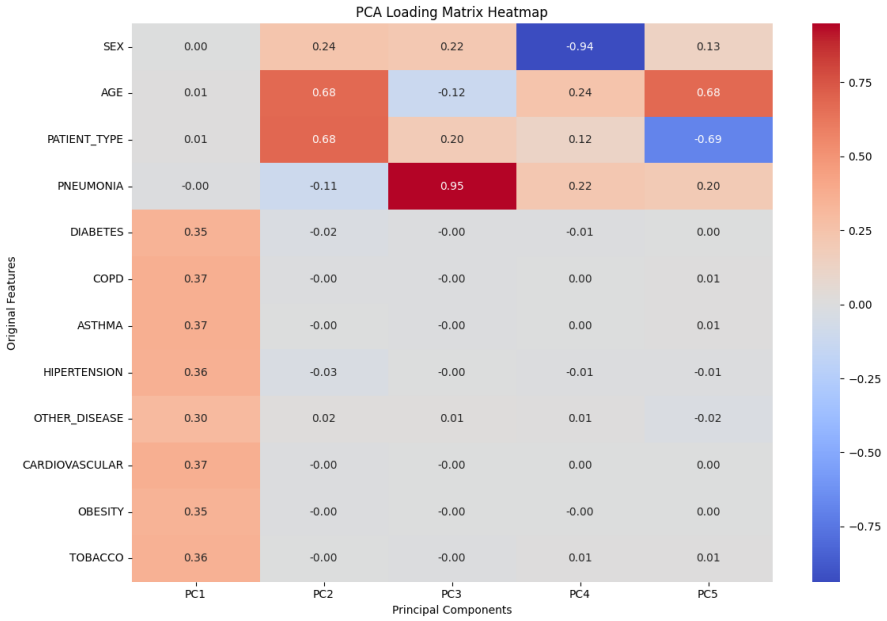

圖 6: PCA 負荷量熱圖

特徵名稱	第一主成分	第二主成分	第三主成分	第四主成分	第五主成分
SEX	0.0046	0.2397	0.2156	-0.9370	0.1340
AGE	0.0082	0.6786	-0.1241	0.2424	0.6816
PATIENT_TYPE	0.0104	0.6850	0.1956	0.1214	-0.6900
PNEUMONIA	-0.0022	-0.1073	0.9485	0.2196	0.2014
DIABETES	0.3477	-0.0190	-0.0022	-0.0060	0.0026
COPD	0.3670	-0.0040	-0.0026	0.0004	0.0075
ASTHMA	0.3696	-0.0000	-0.0018	0.0009	0.0077
HIPERTENSION	0.3613	-0.0253	-0.0013	-0.0085	-0.0072
OTHER_DISEASE	0.3049	0.0184	0.0135	0.0091	-0.0231
CARDIOVASCULAR	0.3690	-0.0020	-0.0010	0.0030	0.0034
OBESITY	0.3489	-0.0042	-0.0004	-0.0008	0.0030
TOBACCO	0.3552	-0.0007	-0.0020	0.0073	0.0062

表 5: PCA 負荷量表

佈在主成分空間中具有一定的區域性特徵，特別是在第二主成分上。

3.4 分析結論

1. **第一主成分的主要貢獻來源**第一主成分捕捉了數據中超過 54% 的變異，主要由慢性疾病相關特徵（如糖尿病、慢性阻塞性肺病、哮喘等）驅動，這些特徵在健康狀況中具有重要影響力。

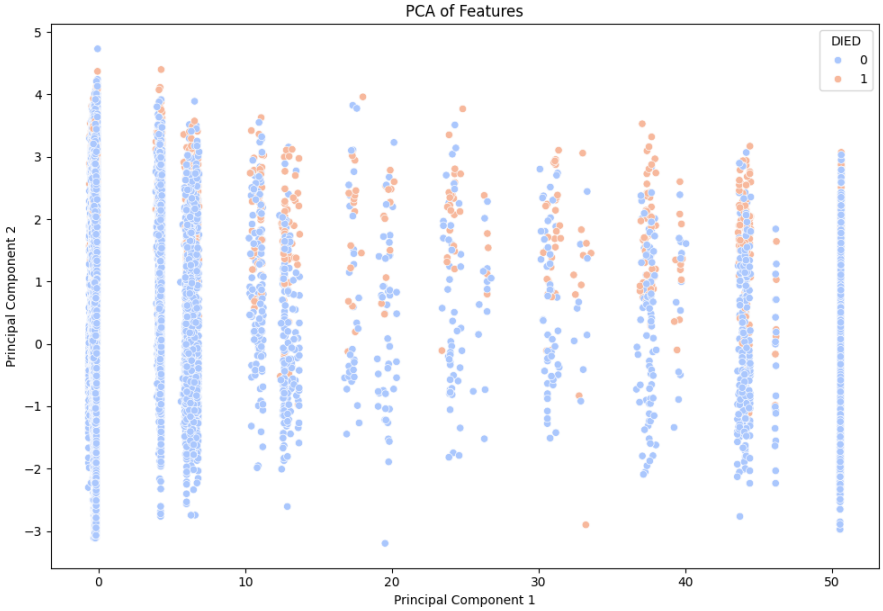

圖 7: PCA 點散圖 (前兩個主成分)

2. **第二主成分的影響因素**第二主成分與年齡（AGE）和患者類型（PATIENT_TYPE）有顯著關聯，表明這些特徵對數據中某些趨勢的分佈具有重要作用。
3. **特徵對死亡的影響** PCA 點散圖（圖 7）顯示，死亡患者（DIED=1）在第二主成分上的分佈與未死亡患者（DIED=0）存在一定差異，可能表明年齡和患者類型在死亡風險中具有區分性。
4. **主成分負荷量的顯示結果**
   - 第一主成分高度依賴於慢性疾病特徵（如糖尿病、COPD、ASTHMA等）。
   - 第二主成分主要由年齡和患者類型。
   - 第四主成分與肺炎（PNEUMONIA）特徵高度相關。

4 主成分分析 (PCA) 後的模型評估與結果

4.1 交叉驗證與模型評估

在進行主成分分析（PCA）後，對篩選出的主要特徵進行了交叉驗證，並使用隨機森林與 XGBoost 模型進行分類評估。以下是交叉驗證的結果：

- **交叉驗證準確度**：
  - 交叉驗證準確度分數：

$$[0.9297, 0.9300, 0.9301, 0.9301, 0.9308]$$

  - 平均準確度：

$$\text{Mean Cross-Validated Accuracy} = 0.9301$$

  - 測試準確度：

$$\text{Test Accuracy} = 0.93$$

4.2 隨機森林模型結果

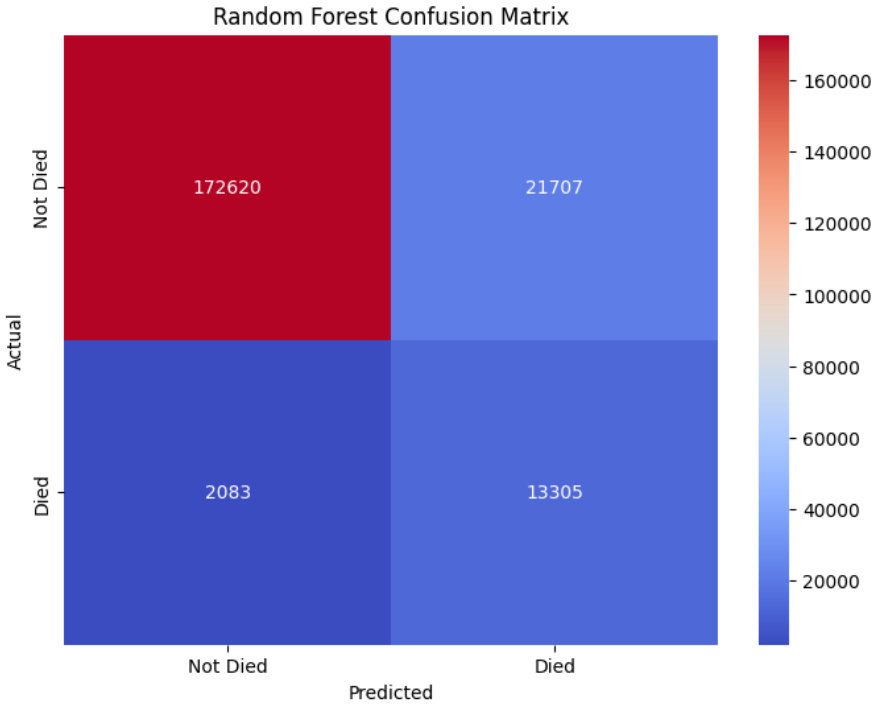

圖 8: 隨機森林模型混淆矩陣

隨機森林的分類報告如下：

```
Precision: 0.99 (Not Died), 0.38 (Died)
Recall:    0.89 (Not Died), 0.86 (Died)
F1-Score:  0.94 (Not Died), 0.53 (Died)
Weighted Avg Accuracy: 0.91
```

如圖 8 所示，隨機森林模型的結果分析如下：

- **未死亡樣本（Not Died）**：
  - 精確率高達 99%，召回率為 89%，表明該模型能準確預測「未死亡」樣本，錯誤分類率較低。
- **死亡樣本（Died）**：
  - 精確率僅為 38%，召回率為 86%，說明該模型能捕捉大部分「死亡」樣本，但錯誤分類為「未死亡」的情況較多。
  - F1 分數為 0.53，顯示該模型在少數樣本（Died）上的表現仍有改進空間。

4.3 XGBoost 模型結果

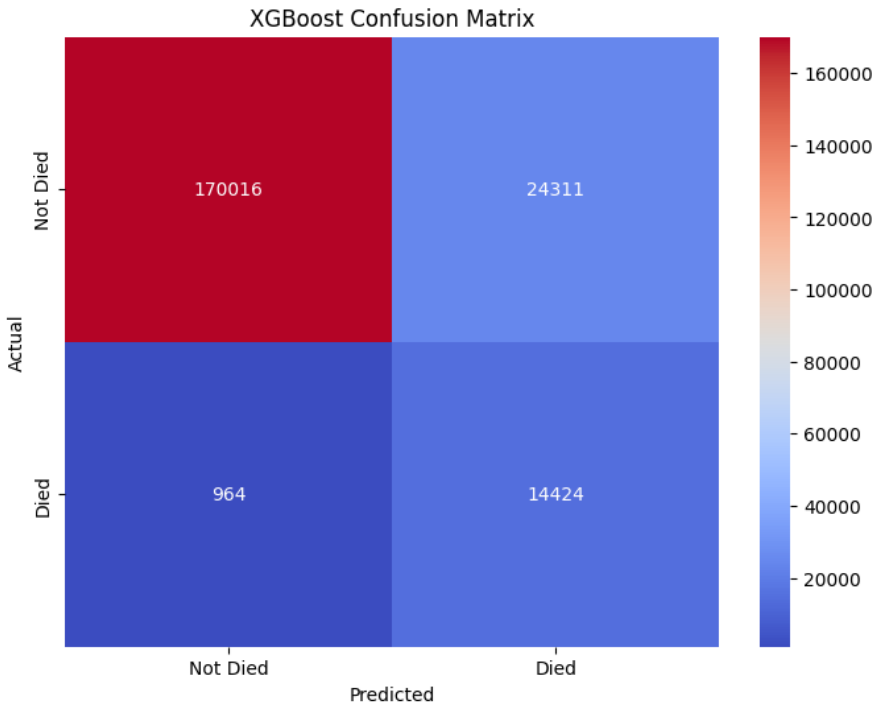

圖 9: XGBoost 模型混淆矩陣

XGBoost 的分類報告如下：

```
Precision: 0.99 (Not Died), 0.37 (Died)
Recall:    0.87 (Not Died), 0.94 (Died)
F1-Score:  0.93 (Not Died), 0.53 (Died)
```

```
Weighted Avg Accuracy: 0.90
```

如圖 9 所示，XGBoost 模型的結果分析如下：

- **未死亡樣本（Not Died）**：
  - 精確率為 99%，召回率為 87%，略低於隨機森林模型，但仍表現出色。
- **死亡樣本（Died）**：
  - 精確率為 37%，召回率提升至 94%，顯示該模型能更好地捕捉「死亡」樣本，但精確率下降，存在一定的誤分類問題。
  - F1 分數與隨機森林相同，均為 0.53，表明該模型對少數樣本的分類能力存在相似的限制。

4.4 模型問題分析與改進

根據模型的分類報告與混淆矩陣，我們發現以下問題：

- **死亡樣本召回率偏低**：隨機森林與 XGBoost 模型對於未死亡樣本（Not Died）具有高準確度與 F1 分數，但對死亡樣本（Died）的召回率明顯偏低（0.86 與 0.94），顯示模型在處理少數樣本時存在偏倚。
- **數據不平衡影響**：數據集中未死亡樣本數量遠大於死亡樣本，導致模型對未死亡樣本的預測表現優異，但難以有效捕捉死亡樣本的特徵。
- **部分變數的重要性不足**：PCA 提取的主成分雖保留了大部分數據的變異性，但並未完全關注與死亡樣本高度相關的變數，可能導致模型難以專注於高風險特徵。

基於上述問題，採用 Cramér's V 方法對變數進行篩選，並重新訓練模型以期提升對死亡樣本的預測能力。

5 基於 Cramér's V 的變數篩選與分析

5.1 變數篩選

根據 Cramér's V 值對各變數與死亡情況（DIED）的相關性檢定結果，我們將變數分為三類：

- **高相關變數**：
  - PATIENT_TYPE (0.5156)
  - PNEUMONIA (0.4694)

  - DIABETES (0.2185)
  - HIPERTENSION (0.2064)
  - RENAL_CHRONIC (0.1235)
- **中低相關變數**：
  - COPD (0.0959)
  - CARDIOVASCULAR (0.0840)
  - PREGNANT (0.0830)
  - SEX (0.0814)
- **低相關變數**：
  - ASTHMA (0.0377)
  - TOBACCO (0.0329)

5.2 篩選與改進的邏輯

- **高相關變數**：這些變數與死亡情況之間的相關性顯著，包含患者類型、是否患有肺炎及糖尿病等，可能對模型有重要解釋力。
- **中低相關變數**：這些變數的相關性雖不如高相關變數，但仍可能提供輔助信息，例如慢性阻塞性肺病與性別。
- **低相關變數**：這些變數與死亡情況的相關性較弱，對模型的影響可能有限，因此不納入後續分析。

最終，我們僅保留高相關與中相關的變數進行模型分析，接下來使用這些變數重新進行模型訓練與交叉驗證，以期提升模型對死亡樣本的召回率。

6 高與中相關變數的隨機森林與 XGBoost 結果分析

在篩選出高相關與中相關變數後，模型的結果如下：

6.1 隨機森林模型結果

- **交叉驗證準確率 (CV Accuracy)**：0.9231
- **分類報告 (Classification Report)**：
  - **未死亡 (Not Died)**：

    * 精確率 (Precision)：0.94
    * 召回率 (Recall)：0.99
    * F1 分數：0.96
  - **死亡 (Died)**：
    * 精確率 (Precision)：0.55
    * 召回率 (Recall)：0.18
    * F1 分數：0.28
  - **整體準確率 (Accuracy)**：0.93
  - Macro Avg：
    * 召回率：0.59
    * F1 分數：0.62
- **混淆矩陣 (Confusion Matrix)**：如圖 10 所示。

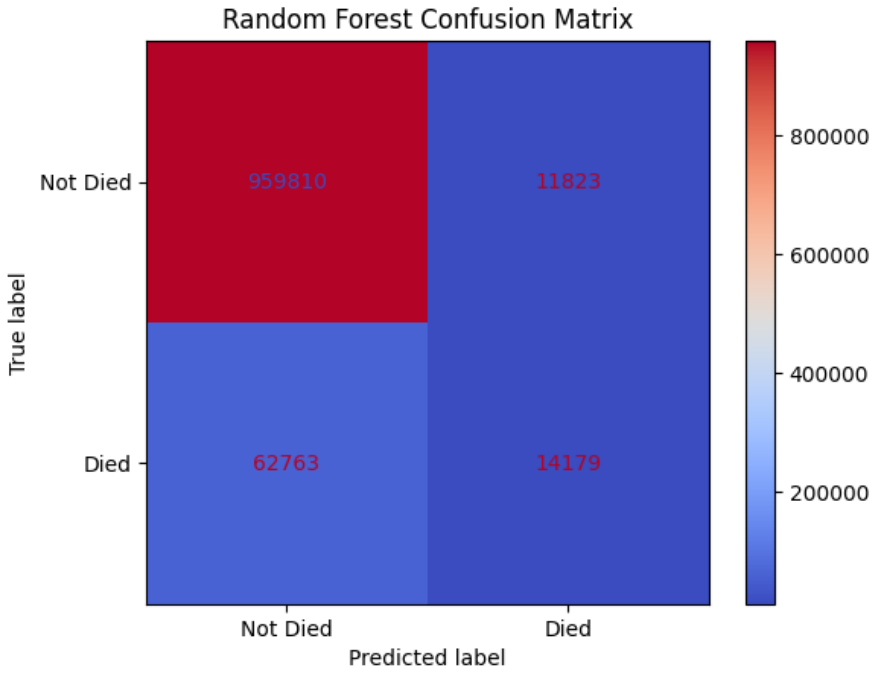

圖 10: 中高變數隨機森林模型的混淆矩陣

6.2 XGBoost 模型結果

- **交叉驗證準確率 (CV Accuracy)**：0.9226
- **分類報告 (Classification Report)**：
  - **未死亡 (Not Died)**：
    * 精確率 (Precision)：0.94
    * 召回率 (Recall)：0.99
    * F1 分數：0.96
  - **死亡 (Died)**：
    * 精確率 (Precision)：0.54
    * 召回率 (Recall)：0.18
    * F1 分數：0.27
  - **整體準確率 (Accuracy)**：0.93
  - 宏平均 (Macro Avg)：
    * 召回率：0.58
    * F1 分數：0.62
- **混淆矩陣 (Confusion Matrix)**：如圖 11 所示。

6.3 結果分析

從上述結果可以觀察到：

- **整體準確率高，但對死亡樣本的預測效果不佳**：
  - 雖然隨機森林和 XGBoost 的整體準確率都達到了 93%，但對於死亡樣本 (Died) 的召回率都只有 18%。
  - 這表明模型在平衡數據集的預測時，對於少數類別（死亡樣本）的表現相對較弱。
- **混淆矩陣顯示錯誤分類集中於死亡樣本**：
  - 大量死亡樣本被錯誤分類為未死亡，特別是在 XGBoost 模型中，錯誤分類數高達 63,087。
  - 這進一步證明模型對於死亡樣本的區分能力需要提升。

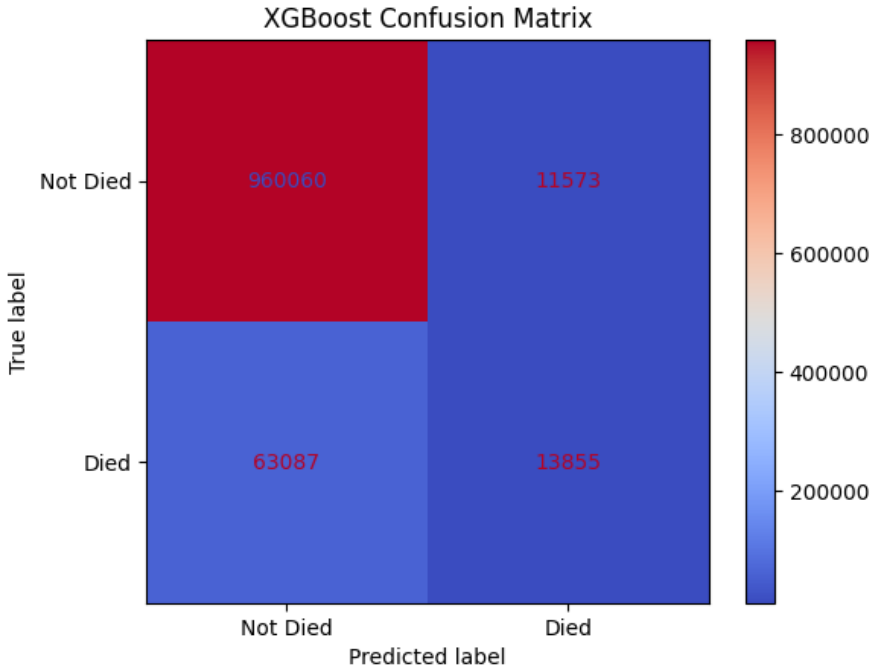

圖 11: 中高變數 XGBoost 模型的混淆矩陣

- **可能原因與改進方向**：
  - **數據不平衡**：死亡樣本數量遠低於未死亡樣本，導致模型傾向於正確分類未死亡樣本。
  - **特徵重要性分佈不均**：高相關變數可能未能完全捕捉死亡風險的異常特徵。
  - **改進建議**：
    * 引入 SMOTE 或其他過採樣方法，平衡數據集。
    * 減少未死亡樣本的權重或調整模型損失函數，以增加對少數類別的敏感性。

7 SMOTE 處理後的模型結果分析

為了改善模型對死亡樣本（DIED=1）的預測能力，我們採用了 SMOTE（Synthetic Minority Oversampling Technique）對數據進行平衡化處理。以下是基於 SMOTE 處理後的隨機森林（Random Forest）與 XGBoost 模型結果。

7.1 隨機森林模型結果

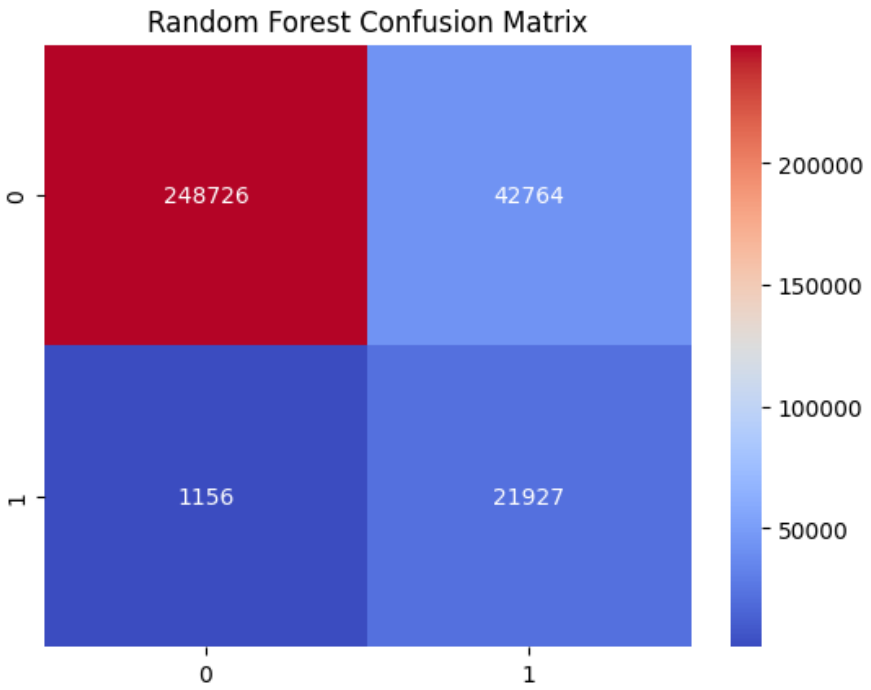

圖 12: SMOTE 處理後隨機森林混淆矩陣

隨機森林分類報告：

	precision	recall	f1-score	support
0	1.00	0.85	0.92	291490
1	0.34	0.95	0.50	23083
accuracy			0.86	314573
macro avg	0.67	0.90	0.71	314573
weighted avg	0.95	0.86	0.89	314573

7.2 XGBoost 模型結果

XGBoost 分類報告：

	precision	recall	f1-score	support
0	1.00	0.85	0.92	291490
1	0.34	0.95	0.50	23083

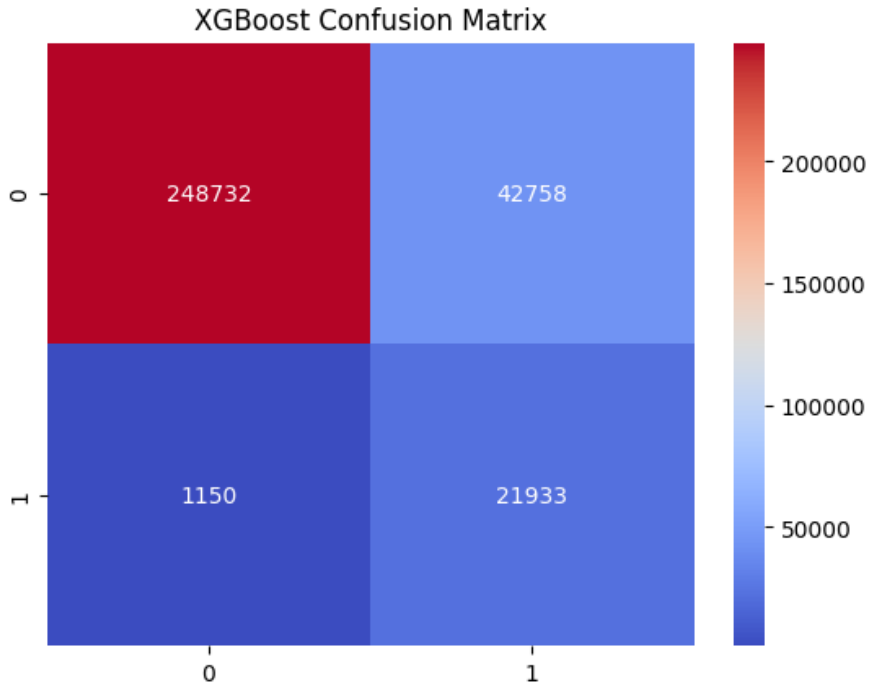

圖 13: SMOTE 處理後 XGBoost 混淆矩陣

accuracy			0.86	314573
macro avg	0.67	0.90	0.71	314573
weighted avg	0.95	0.86	0.89	314573

7.3 分析與結論

SMOTE 處理後，兩個模型對死亡樣本（DIED=1）的召回率（recall）顯著提升至 95%。以下是具體觀察與分析：

- **召回率顯著提升**：相較於未使用 SMOTE 的模型，對死亡樣本的召回率從原本的 18% 提升至 95%，顯示模型在識別死亡樣本時表現大幅改善。
- **精確率下降**：死亡樣本的精確率（precision）下降至 34%，顯示部分未死亡樣本被錯誤分類為死亡樣本，這可能是過採樣帶來的影響。
- **整體模型準確率**：模型整體準確率維持在 86%，表明模型在平衡未死亡與死亡樣本的預測時，表現穩定。
- **適用場景**：在實際應用中，若重視對死亡樣本的召回（例如在公共衛生或危急情況管理中），該模型提供了一個可行解決方案。

綜合來看，SMOTE 處理有效改善了模型對小樣本（死亡樣本）的識別能力，但也引入分誤分類問題。在需要高召回率的場景下，未來可以通過調整分類閾值或結合其他技術（例如異常檢測）進一步優化模型表現。

8 總結與模型解釋

8.1 模型解釋

最終模型的特徵係數如下：根據表 6最終模型的主要依賴特徵如下：

表 6: 最終模型特徵係數

特徵名稱	係數
PATIENT_TYPE	1.637
SEX	0.185
COPD	0.065
CARDIOVASCULAR	0.040
RENAL_CHRONIC	0.031
DIABETES	-0.018
HIPERTENSION	-0.069
PNEUMONIA	-0.174

- **PATIENT_TYPE（患者類型）**：係數最高（1.637），表明住院患者的死亡風險顯著高於非住院患者，由於住院患者通常病情更為嚴重。該變數是影響死亡風險的最重要指標。
- **COPD（慢性阻塞性肺病）**、**CARDIOVASCULAR（心血管疾病）**、**RENAL_CHRONIC（慢性腎病）**：這些特徵的係數為正，顯示這些慢性疾病會增加死亡風險，符合臨床觀察結果。
- **DIABETES（糖尿病）**：係數為負（-0.018），這可能是由於糖尿病與其他共病（如高血壓或腎病）高度相關，在模型中控制其他變數後，影響被削弱。可能反映了部分糖尿病患者接受了更積極的治療干預。
- **HIPERTENSION（高血壓**與 **PNEUMONIA（肺炎）**：這些特徵的負係數（-0.069 和 -0.174）表明這些患者接受了針對性治療，導致其死亡風險相對降低。負係數不代表這些特徵本身降低風險，可能在模型中被其他高風險特徵所覆蓋。

8.2 模型性能與改進後效果

- **基本性能**：原始模型在死亡樣本上的召回率偏低，特別是隨機森林與 XGBoost 模型，對死亡樣本的識別能力不佳。
- **改進措施**：通過 SMOTE 平衡數據，顯著提升了模型對死亡樣本的召回率（95%），但同時引入了精確率下降（34%）的問題。
- **適用場景**：最終模型更適合高敏感性的應用場景，例如公共健康監測或重症患者管理。需要注意的是，假陽性的增加可能對資源分配產生影響。

8.3 結論與未來改進方向

結論：

- **關鍵特徵**：患者類型（PATIENT_TYPE）是最具解釋力的變數，與其他慢性疾病（COPD、CARDIOVASCULAR 等）共同決定了死亡風險。
- **模型改進效果**：通過 SMOTE 平衡數據，顯著提高了對死亡樣本的識別能力，但模型在精確率上有所犧牲。

未來改進方向：

1. **特徵工程**：基於現有變數進一步生成組合特徵（例如多重慢性病指標），提升模型的穩定性與解釋力。
2. **結合異常檢測**：使用異常檢測技術（如 Isolation Forest）進一步識別死亡高風險群體，避免過度依賴分類模型。
3. **多模型融合**：結合其他算法（如支持向量機或神經網絡）進行集成建模，提升模型的整體表現，特別是在小樣本類別的識別上。

9 模型挑戰

在平衡數據後，模型的召回率提升（如死亡樣本召回率達 95%），但精確率下降至 34%。為應對此挑戰，我們採取以下方法：

- 調整分類值，根據場景需求在召回率和精確率間取得平衡。
- 根據應用需求，優先選擇高召回率或高精確率模型。

9.1 結論

共線性問題和召回率-精確率平衡是模型分析的挑戰。我們通過特徵工程和數據處理緩解了這些問題，為後續應用提供了可行方案。

10 心得

經過這份報告讓我學習更多處理資料偏差的方法，更了解分類模型的作用與建模所要注意的地方。其中對於大數據的處理有更多方法的加入，不是一個步驟就到位，需要根據不同的方法篩選變數的部分最後才能達到想要的結果，並且對於缺失值不是將其刪除就好，有更多其他的方法可以處理缺失值，對往後的論文處理可以有更多的選擇。