機器學習與大數據分析技術

Final Project

羅勻瑄、黃世君、劉定睿

日期：2026-06-03

目錄

第 1 章｜研究動機與問題定義

1.1 研究背景

美國鴉片類藥物危機的宏觀脈絡、Connecticut 州的嚴重程度、為什麼需要用資料分析來理解死亡模式。

1.2 研究目標

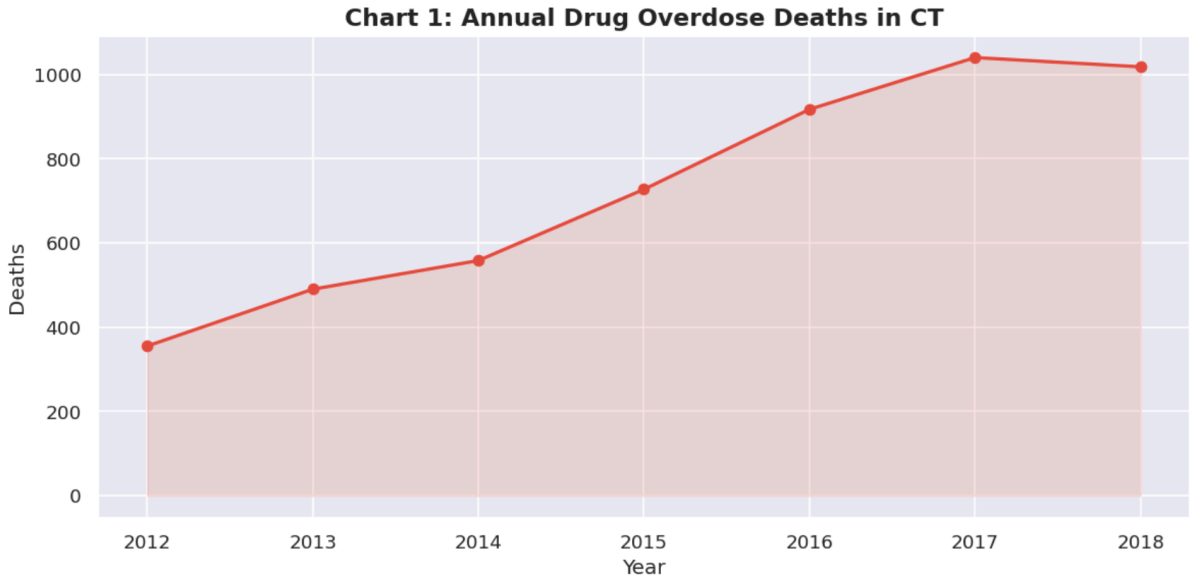
- 死亡人數的年度趨勢為何？
- 主要致死藥物是否改變？
- 死亡案例能否被自動分成有意義的族群？
- 人口學特徵能否預測多重用藥致死？
- 藥物之間有哪些常見的致命搭配？
- 是否存在特別極端的死亡模式？

1.3 研究方法

本研究使用機器學習與資料分析方法進行數據分析，運用原理包含資料概況分析、分群分析、監督式學習、非監督式學習進行分析。

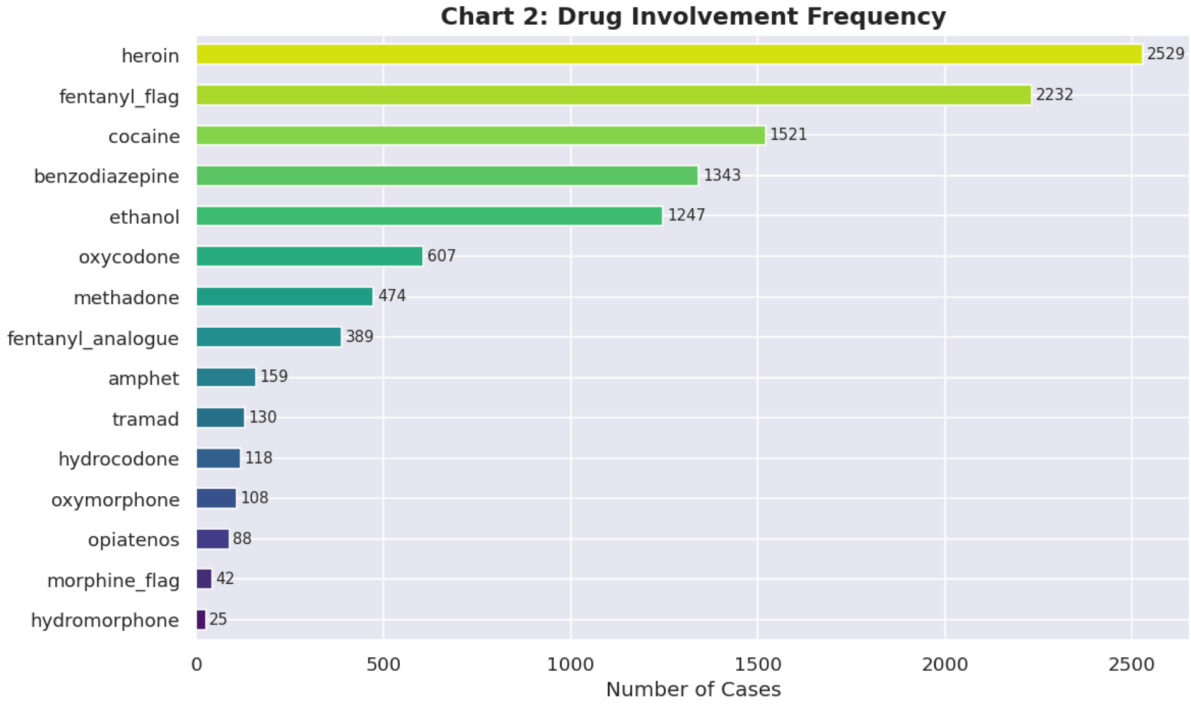
第 2 章｜EDA — 資料概況與發現

資料集共 5,105 筆死亡記錄、42 個欄位，涵蓋 2012–2018 年 Connecticut 的意外藥物過量死亡案例。年度死亡人數從 2012 年的約 350 人激增至 2017 年的 1,040 人高峰（近 3 倍成長），2018 年首次微降至約 1,020 人。

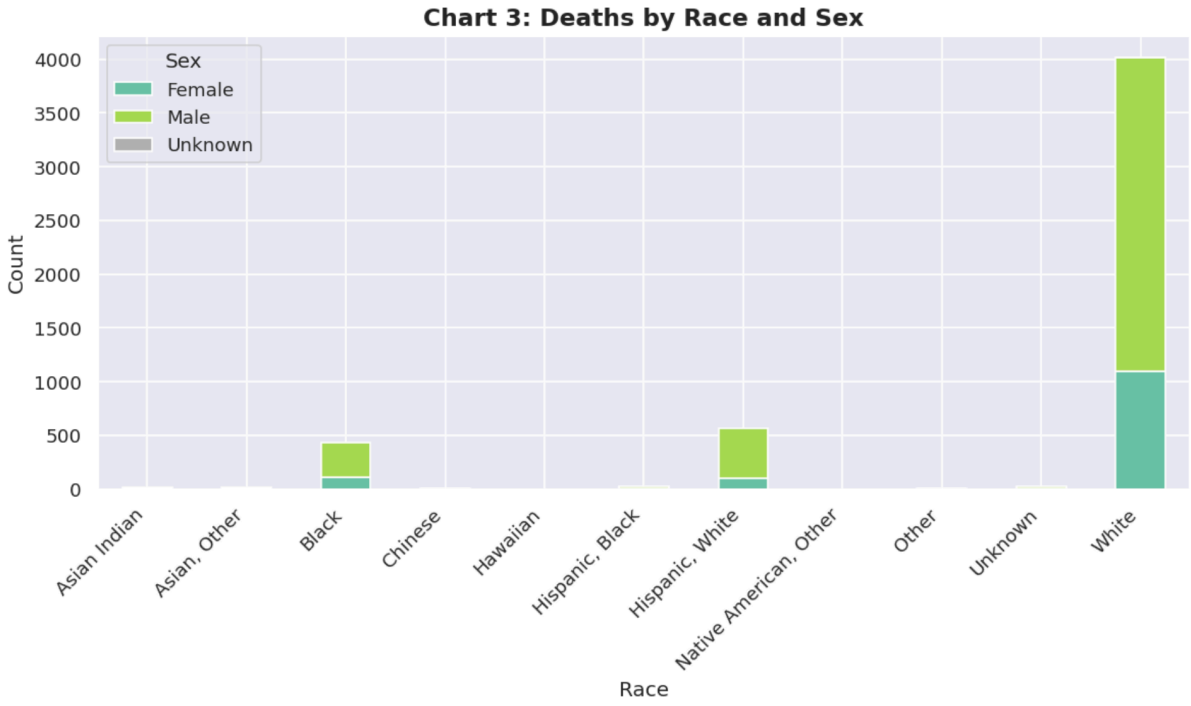


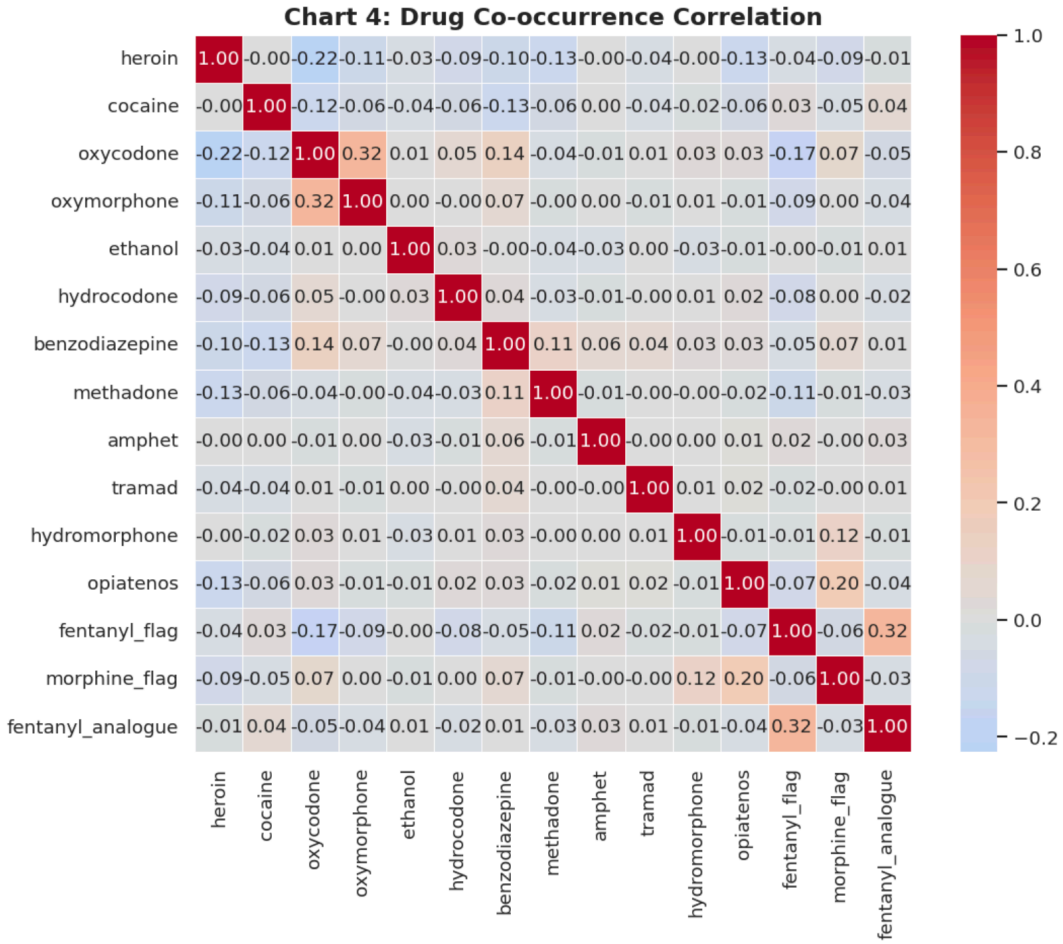
藥物涉及頻率前五名依序為：

- heroin（2,529 例，49.5%）
- fentanyl（2,232 例，43.7%）
- cocaine（1,521 例，29.8%）
- benzodiazepine（1,343 例，26.3%）
- ethanol（1,247 例，24.4%）



人口學方面，死者以白人男性為絕對多數（White 佔 78.4%，Male 佔 74.0%），中位年齡 42 歲。





- 整體觀察：相關性普遍極低
  全圖除了對角線（自己跟自己 = 1.0）之外，幾乎所有數字都在 `-0.20` 到 `+0.32` 之間，大多數接近 0。這代表這 15 種藥物在死亡案例中的共現關係非常鬆散，沒有任何一對藥物有「強相關」。
  這呼應了 PCA 的結論：每種藥物帶來的資訊幾乎是獨立的，不能互相替代。

- 值得注意的正相關（淺紅色格子）

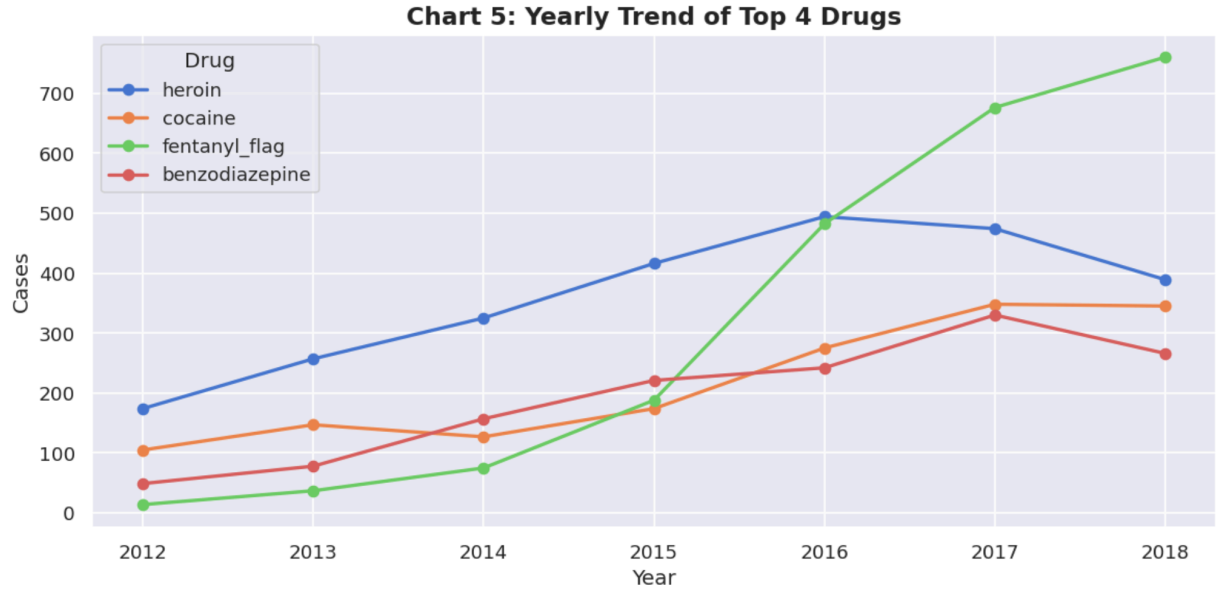
  - oxycodone ↔ oxymorphone（r = 0.32）— 最強的正相關
    這兩種都是半合成鴉片類止痛藥，oxymorphone 本來就是 oxycodone 的代謝產物之一，化學結構相似、藥理機轉相同，都源自醫療處方。
    用 oxycodone 的人同時用 oxymorphone 的機率顯著高於其他組合，反映的是「處方止痛藥多重使用」的模式。

  - fentanyl_flag ↔ fentanyl_analogue（r = 0.32）— 並列最強
    芬太尼本體跟芬太尼類似物同時出現的比率是所有組合裡最高的。
    白話說就是：如果驗出了芬太尼類似物，幾乎一定也會驗出芬太尼本體。

這個結果與 Apriori 關聯規則的 Confidence=100% 完全一致，在化學上也合理——類似物是在芬太尼結構上修改而來的，通常是製造者將兩者混合或類似物本身就含有芬太尼前驅物。

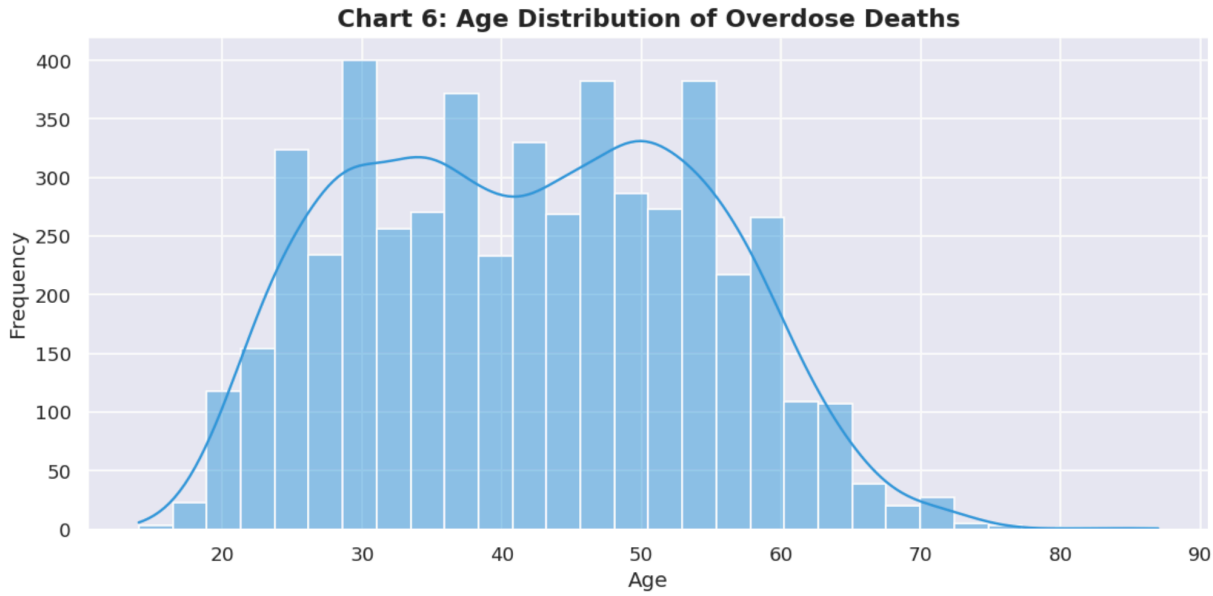
- opiatenos ↔ morphine_flag（r = 0.20）
  「非特定鴉片類」與「非海洛因來源嗎啡」同時出現。這兩個欄位本來就有概念上的重疊，當法醫無法確定鴉片類來源時，這兩欄都可能被標記，部分解釋了相關性。

- oxycodone ↔ benzodiazepine（r = 0.14）
  處方止痛藥與安眠鎮定藥同時出現。
  這反映了臨床上常見的「止痛藥 + 安眠藥」雙重處方問題，兩者都是醫師可開立的合法藥物，但合併使用會大幅抑制呼吸中樞，是處方藥過量死亡的典型路徑。



逐年趨勢圖結構性變化：

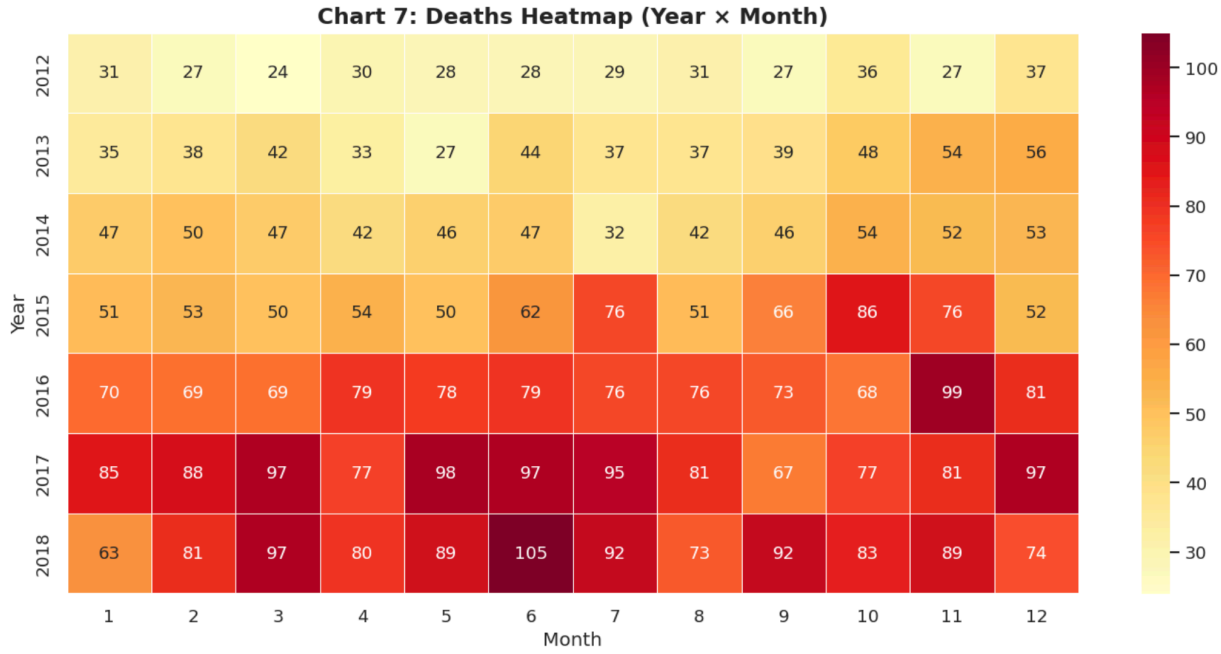
- fentanyl 從 2012 年僅 14 例爆炸性成長至 2018 年的 752 例，於 2016 年超越 heroin 成為第一大致死藥物，
- heroin 則在 2017 年後開始下降。這反應了美國鴉片類危機從處方藥→海洛因→合成類鴉片的典型演變路徑。



年齡分布呈現明顯的雙峰結構（Bimodal Distribution），KDE 曲線在 28–30 歲和 48–54 歲各有一個波峰，兩峰之間在 40 歲附近出現凹陷。這代表藥物過量死亡並非均勻分布在各年齡層，而是集中在兩個截然不同的族群：

1. 第一峰（25–35 歲，峰值約 400 人）：年輕成癮者，通常是較早接觸海洛因或芬太尼等街頭毒品的族群，用藥史較短但劑量控制能力差，容易因初期耐受性不足而過量致死。
2. 第二峰（45–55 歲，峰值約 330 人）：中年重度用藥者，許多人可能從合法的處方止痛藥（oxycodone、hydrocodone）開始，經過多年逐步發展為多重用藥。這個年齡層同時也是 methadone 替代療法和 benzodiazepine 處方的主要使用族群，混用風險更高。

分布的右尾在 60 歲以上快速下降（65 歲以上僅約 110 人，70 歲以上不到 40 人），左尾在 20 歲以下也極少（不到 25 人），最年輕的死者為 14 歲，最年長為 87 歲。整體中位數為 42 歲，落在兩峰之間的凹陷處，**這意味著中位數本身並不代表最常見的死亡年齡，而是被兩個峰「拉平」的結果**。



年份間的結構性成長與月份間的季節性差異
年份維度：顏色從淡黃逐年加深至深紅

- 2012 年全年每月死亡僅 24–37 人
- 2017–2018 年大多數月份超過 80 人
- 這個「整排變深」的趨勢不是任何單一月份造成的，而是所有月份同步上升，反映芬太尼危機是全年性的結構性惡化，不受季節影響。

整份資料的極值：

- 最高：2018 年 6 月 = 105 人，是整張圖中唯一突破 100 的格子，顏色最深
- 最低：2012 年 3 月 = 24 人，顏色最淡

月份維度（由左往右讀）：沒有明顯的季節性規律。不同年份的高峰月份都不一樣

- 2015 年是 10 月（86 人）
- 2016 年是 11 月（99 人）
- 2017 年是 5 月（98 人）
- 2018 年是 6 月（105 人）

如果有穩定的季節性，應該每年同一個月份都偏高，但這張圖顯示高峰月份在各年間隨機分散。這代表藥物過量死亡的驅動力是年度趨勢（芬太尼供應量增加），而非季節性因素（如冬季憂鬱或夏季社交活動）。

一個值得注意的反直覺現象：2018 年 1 月（63 人）明顯低於 2017 年同月（85 人）。2018 年的整體死亡數微降（1,012 vs 1,040），而降幅主要集中在年初（1 月和 2 月），下半年反而持平甚至升高（9 月 92 人 > 2017 年 9 月 67 人）。這暗示 2018 年

的「微降」可能不是持續性的改善，而是年初的短暫波動。

特徵工程

從原始 42 欄成功擴展至 62 欄（Label Encoded 版本），新增了：

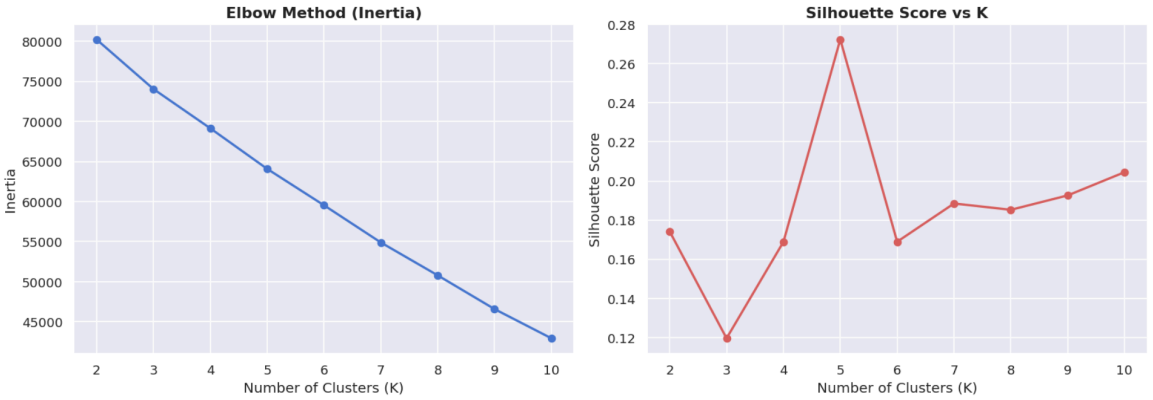
- 5 個時間特徵（year, month, day_of_week, quarter, is_weekend）
- 4 個衍生特徵（drug_count, poly_drug, any_opioid_derived, any_stimulant）
- 6 個 encoding 欄位。

15 個藥物二元欄位全部統一為 0/1 格式。Min-Max 和 Z-score 兩套縮放均正常（mean≈0, std≈1 for Z-score）。

第 3 章｜分群分析

分群分析用於輔助分析相似的死亡模式

```
分群特徵矩陣 Shape: (5105, 17)
使用特徵: ['age', 'heroin', 'cocaine', 'oxycodone', 'oxymorphone', 'ethanol', 'hydrocodone', 'benzodiazepine', 'methadone', 'amphet', 'tramad', 'hydromorphone', 'opiatenos', 'fentanyl_flag', 'morphine_flag', 'fentanyl_analogue', 'drug_count']
```



- 左圖：Elbow Method（手肘法）
  Inertia 從 K=2 的 80,000 到 K=10 的 43,000 幾乎呈完美線性下降，完全沒有明顯的 Elbow。代表這份資料在特徵空間中沒有天然、清楚的分群邊界，不管切幾群，群內距離都在穩定地縮小，不存在某個 K 值讓下降速度突然變慢。光靠 Elbow 法無法決定 K。

- 右圖：Silhouette Score
  Silhouette 出現了明顯的起伏：
  - K=3 時分數急跌到 0.12（最低），代表 3 群的切法非常糟糕，很多點被分到了錯誤的群
  - K=5 時出現尖峰（0.272），遠高於左右兩側
  - K=6 之後分數下滑到 0.17，之後又緩慢回升

結論：選 K=5 是合理的，它是 Silhouette 唯一的顯著尖峰。
但整體分數都在 0.12–0.27 之間偏低，說明 5 群的邊界是模糊的，不是乾淨的分割。

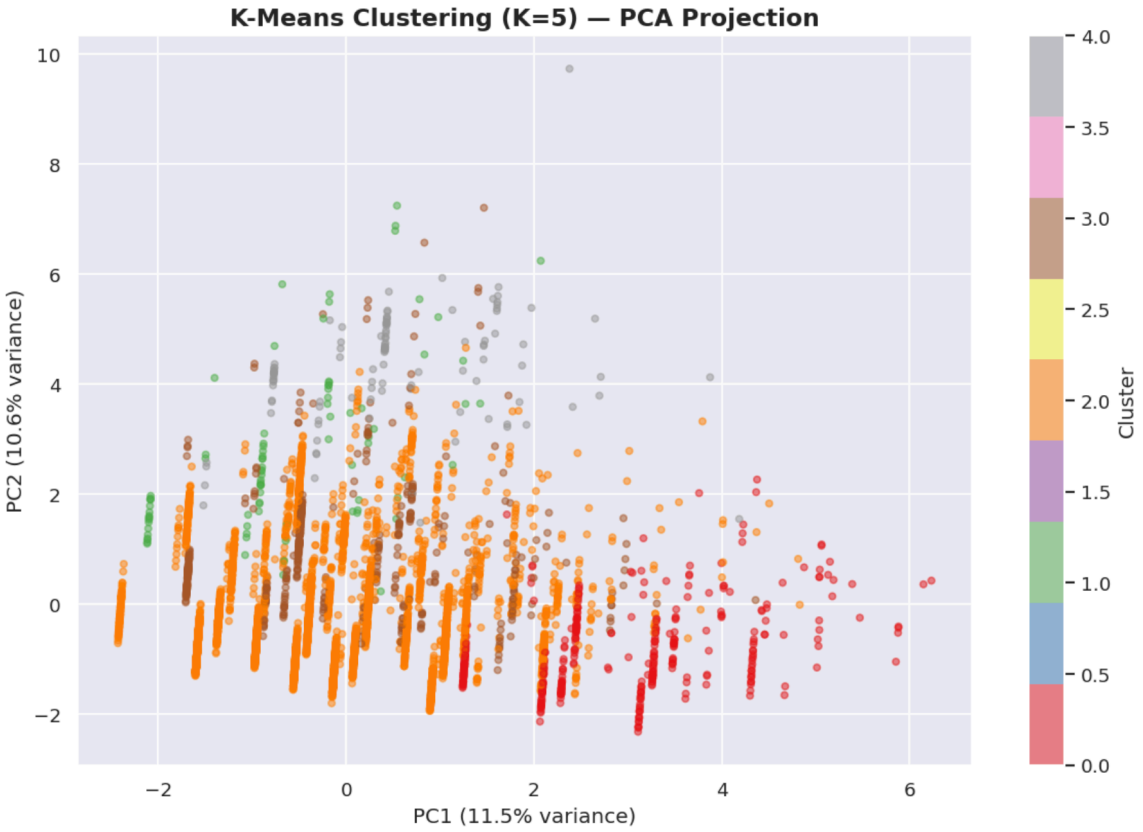
Silhouette Score 在 K=5 時出現明顯尖峰（0.2724），顯著優於其他 K 值。K-Means 的 5 群解釋如下：

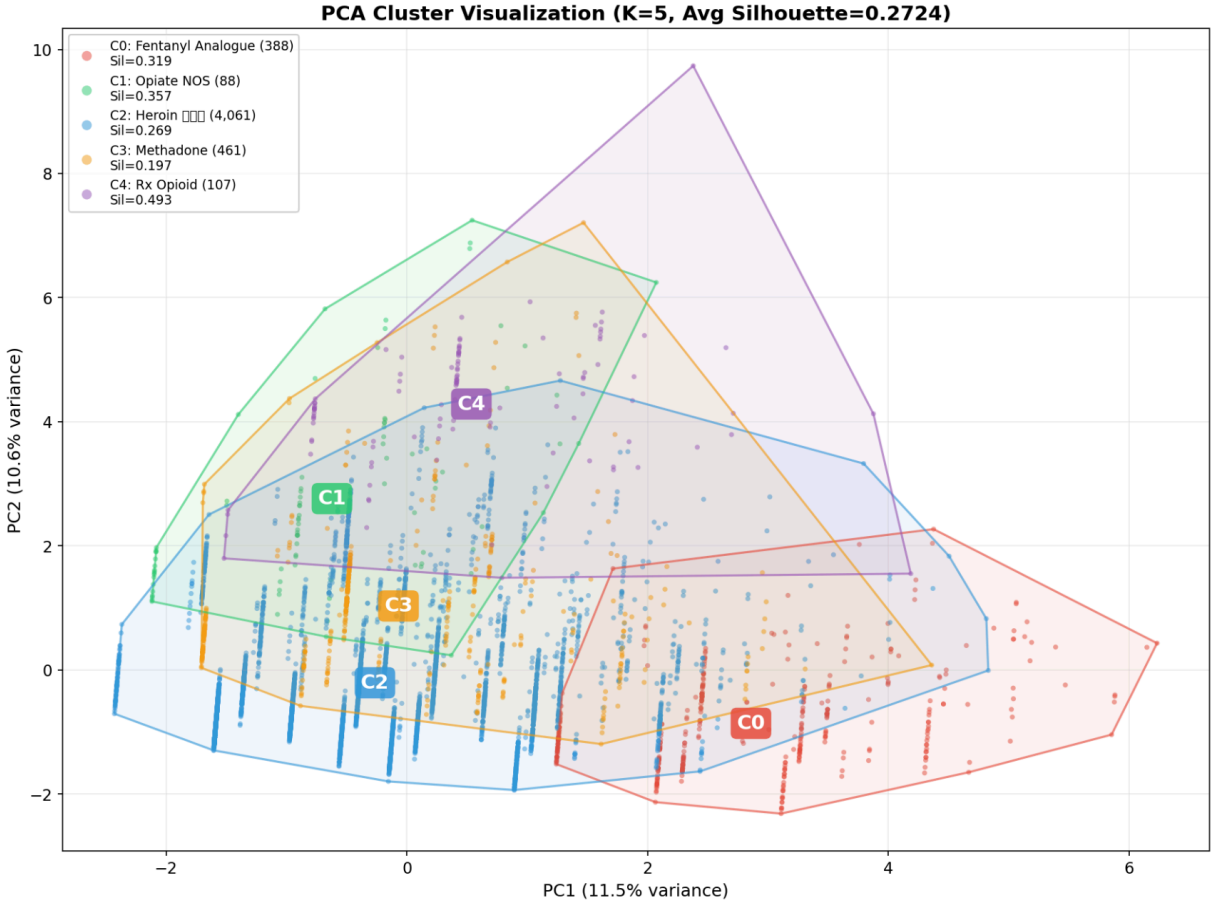
群	人數	主要特徵
Cluster 0 (388人)	fentanyl_analogue=100%, fentanyl=99.7%, drug_count=3.6	**fentanyl 類似物多重用藥群**
Cluster 1 (88人)	opiatenos=100%, heroin=0%, fentanyl=17%	**非特定鴉片類藥物群**
Cluster 2 (4,061人)	最大群，drug_count=1.96，heroin=54.4%	**典型用藥群（heroin 為主）**
Cluster 3 (461人)	methadone=94.4%, benzo=43%	**美沙酮治療相關死亡群**
Cluster 4 (107人)	oxymorphone=100%, oxycodone=82.2%, drug_count=3.1	**處方類鴉片多重用藥群**

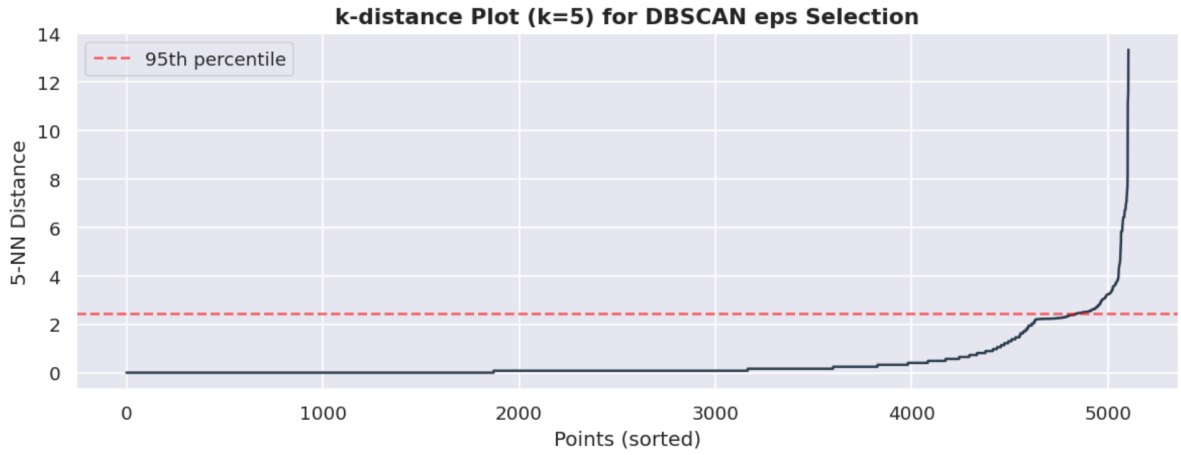
三種方法的 Silhouette 比較：

- Hierarchical 最佳（0.298）
- K-Means 次之（0.272）
- DBSCAN 最低（0.243）

整體 Silhouette 偏低（<0.3）是合理的——藥物過量死亡的用藥組合本身就高度重疊，不存在涇渭分明的分群邊界。


K-Means Clustering (K=5) — PCA Projection
PC2 (10.6% variance)
PC1 (11.5% variance)
Cluster
10
8
6
4
2
0
−2
−2
0
2
4
6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

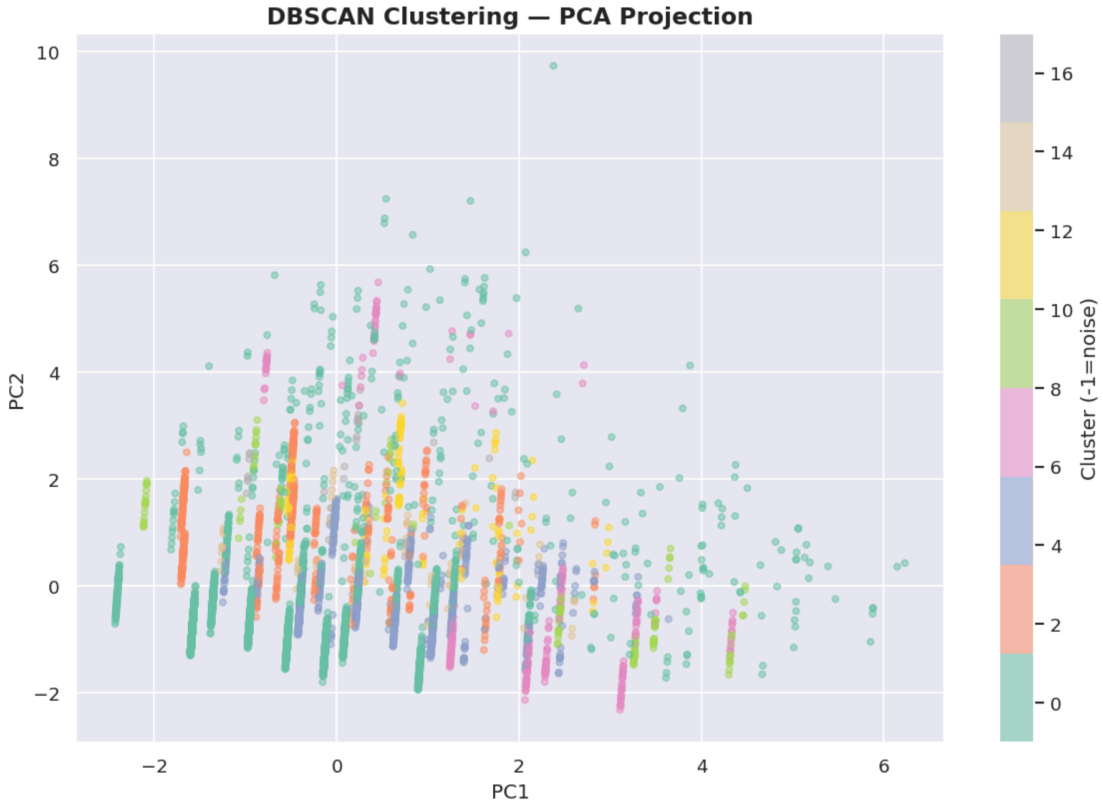

PCA Cluster Visualization (K=5, Avg Silhouette=0.2724)
C0: Fentanyl Analogue (388)
Sil=0.319
C1: Opiate NOS (88)
Sil=0.357
C2: Heroin □□□ (4,061)
Sil=0.269
C3: Methadone (461)
Sil=0.197
C4: Rx Opioid (107)
Sil=0.493
C0
C1
C2
C3
C4
PC2 (10.6% variance)
PC1 (11.5% variance)
10
8
6
4
2
0
−2
−2
0
2
4
6

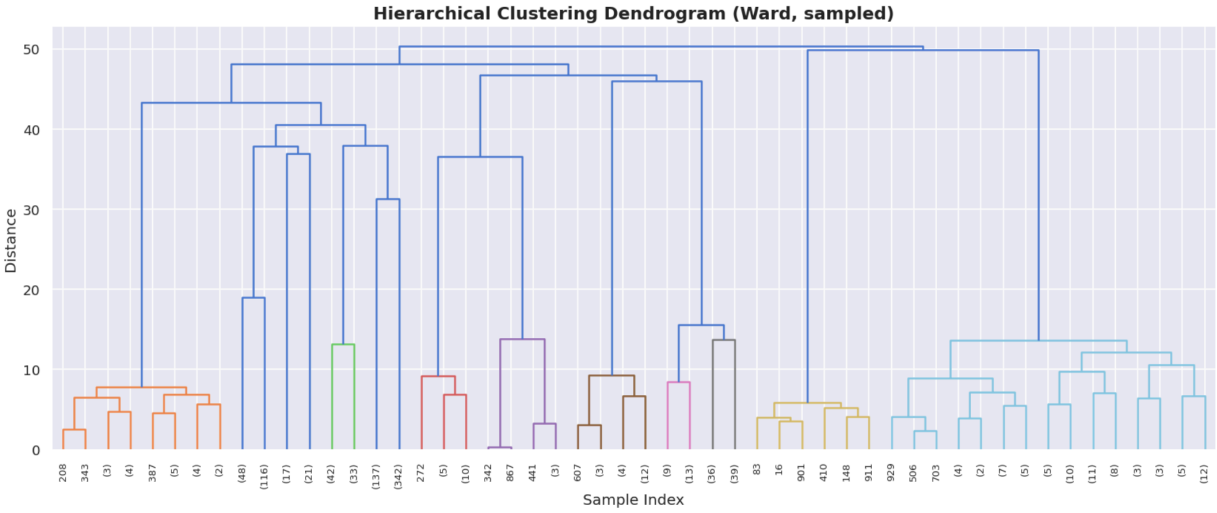


- 0 到 約 1,900 點：距離幾乎為 0：代表大量的死者在特徵空間中完全相同或非常接近（因為多數人只用海洛因，特徵向量幾乎一樣）
- 約 1,900 到 4,500 點：距離緩慢上升
- 約 4,500 點之後：距離急劇上升到 13+：真正的離群點
- 紅色虛線（95th percentile ≈ 2.46）是自動選取的 eps 值
  理論上好的 eps 應該落在曲線開始急劇上升的「肘點」。從圖形看，曲線大約在第 4,500 點附近開始陡增，95th percentile 取到約 2.46，算是合理的選擇。

這張圖的問題：

- 前 1,900 個點的距離幾乎是 0，意味著這份資料有大量的重複特徵向量，代表很多人的藥物組合完全一樣。
- DBSCAN 在這種高度離散的二元資料上，幾乎所有人都會被判定為「密度足夠」的核心點，最終全部被歸入同一個巨大的群，這正是 DBSCAN 在這個資料集上效果不好的根本原因。


DBSCAN Clustering — PCA Projection
PC1
PC2
Cluster (-1=noise)


Hierarchical Clustering Dendrogram (Ward, sampled)
Distance
Sample Index

- 分析分群方法

```
  Calinski-Harabasz Index:
    K-Means (K=5):       451.68
    Hierarchical (K=5):  449.88
    DBSCAN (excl. noise):    391.38

  Davies-Bouldin Index:
    K-Means (K=5):       1.3934
    Hierarchical (K=5):  1.3927
    DBSCAN (excl. noise):    1.4787

  分群穩定性分析 (10 次 Bootstrap):
    Adjusted Rand Index (vs 原始): 0.2560 ± 0.1491
    解讀: ARI > 0.8 = 穩定, 0.5-0.8 = 中等, < 0.5 = 不穩定

=== 分群評估總表 ===
              Silhouette  CH Index  DB Index
K-Means           0.2724  451.6834    1.3934
Hierarchical      0.2980  449.8827    1.3927
DBSCAN            0.2433  391.3772    1.4787

Silhouette: 越大越好 | CH Index: 越大越好 | DB Index: 越小越好
```

第 4 章｜監督式學習

- 目標變數： `poly_drug` （體內驗出 ≥ 2 種藥物 = 1，否則 = 0）
- 特徵： `age` , `year` , `month` , `day_of_week` , `sex_enc` , `race_enc` , `mannerofdeath_enc` , `location_enc`
- 三個模型： `Logistic Regression` 、 `Random Forest` 、 `Hist Gradient Boosting`

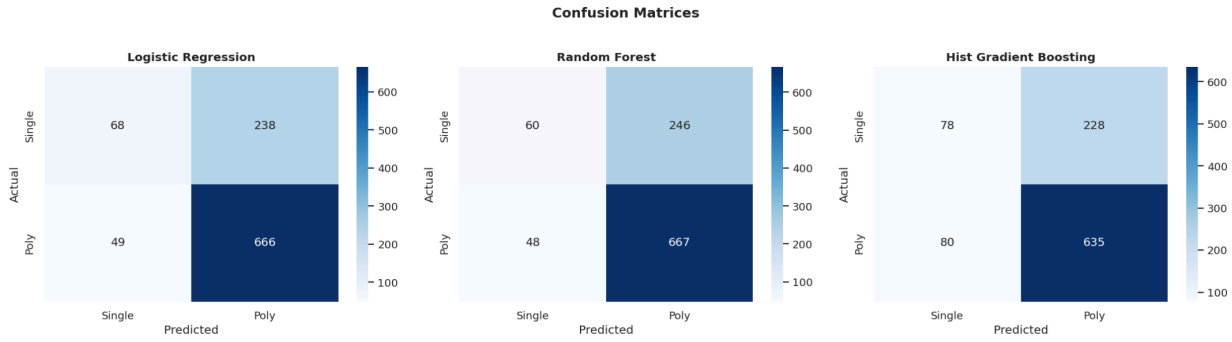
預測多重用藥：

模型	Accuracy	F1	AUC	CV F1
Logistic Regression	**0.719**	**0.823**	0.654	0.819±0.005
Random Forest	0.712	0.819	0.659	0.813±0.006
Hist Gradient Boosting	0.698	0.805	0.630	0.797±0.006

測試集共 1,021 筆，其中：

Single Drug（單一用藥）：306 人（30%）
Poly Drug（多重用藥）：715 人（70%）

- 左上：True Negative (實際 Single，預測 Single) ← 猜對單一用藥
- 右上：False Positive (實際 Single，預測 Poly) ← 誤判成多重用藥
- 左下：False Negative (實際 Poly， 預測 Single) ← 漏掉多重用藥
- 右下：True Positive (實際 Poly， 預測 Poly) ← 猜對多重用藥

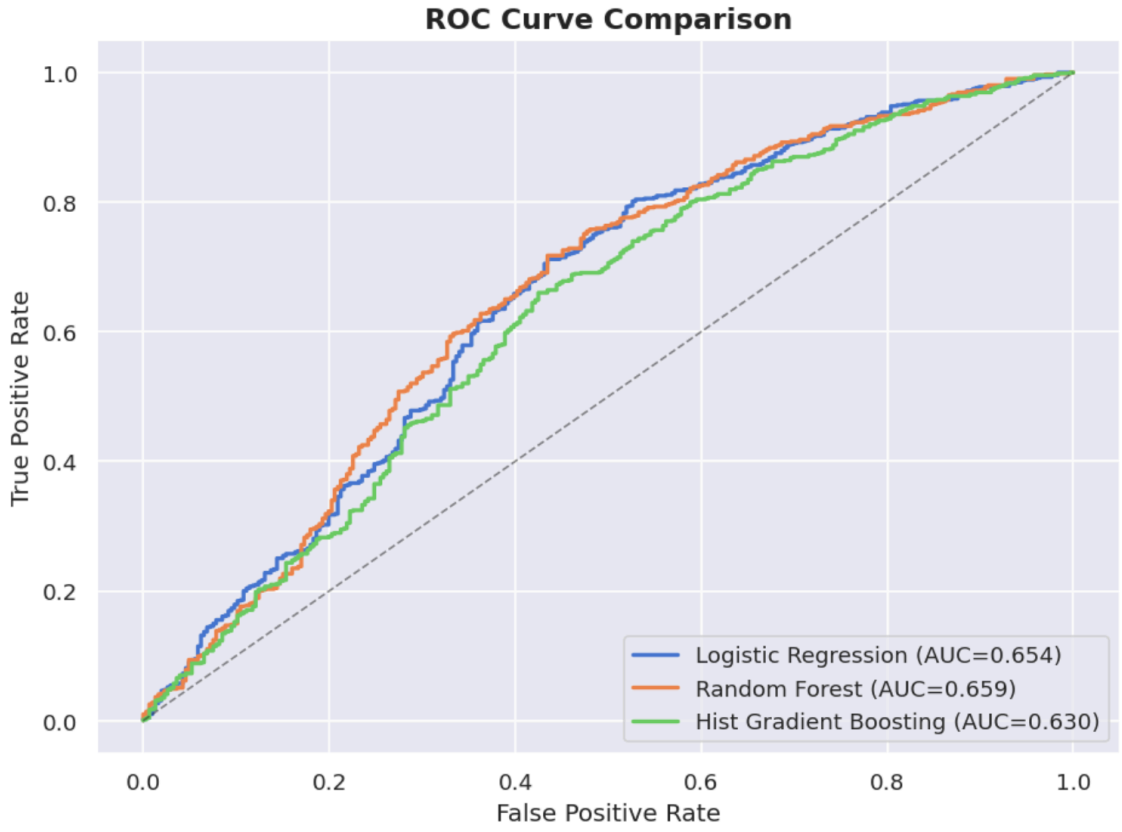


- 三個模型都有同樣的根本問題：
  - 對 Poly Drug 的辨識能力強（89–93%）
  - 對 Single Drug 的辨識能力極弱（20–26%）

    - 右下角深藍色（True Positive = 635–667）遠大於左上角（True Negative = 60–78），矩陣視覺上非常不對稱。

- 這是 70:30 類別不平衡的典型症狀。
  模型學到「猜 Poly 通常不會錯」，因為 70% 的人本來就是多重用藥。

- 比較三個模型的差異：
    - Hist Gradient Boosting 在 Single Drug 的 Recall 稍微好一點（78 vs 60/68），代價是 Poly Drug 的 False Negative 增加（80 vs 49/48）。它在兩類之間取了稍微更平衡的權衡
    - Logistic Regression 和 Random Forest 的矩陣幾乎一模一樣，分別是 668/679 個 True，353/306 個 Error，說明兩者學到的決策邊界非常相似

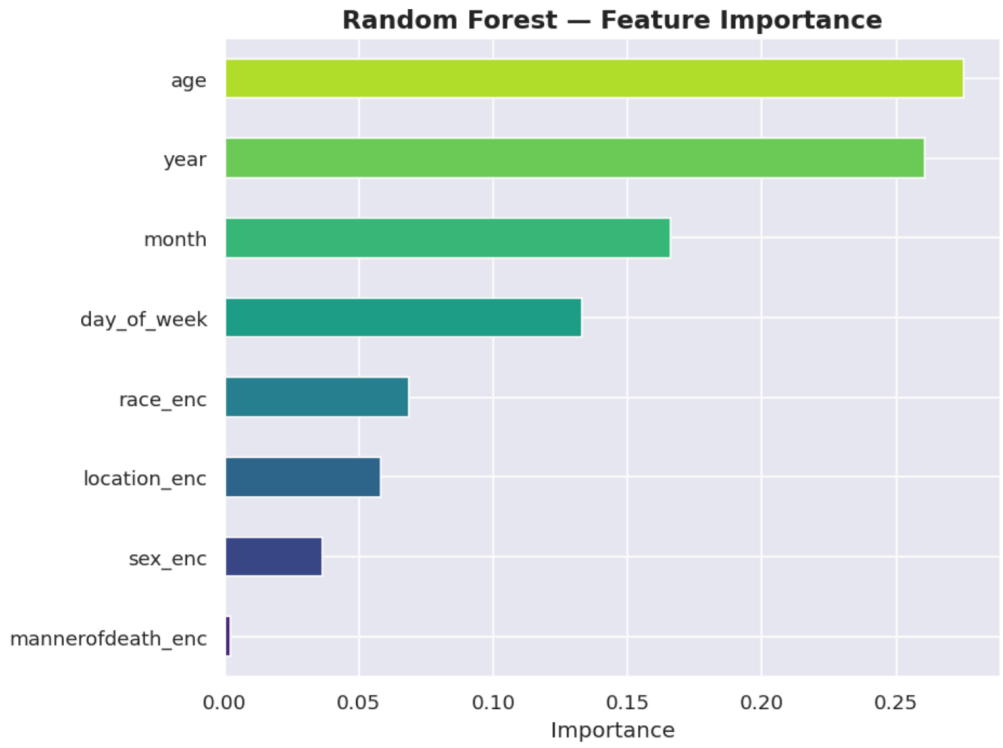


三個模型的 AUC 分別為

- Random Forest 0.659
- Logistic Regression 0.654
- Hist Gradient Boosting 0.630

差距極小（最大僅 0.029），且三條曲線在圖上幾乎重疊，說明不同複雜度的模型在這個問題上都到了同一個天花板。
三個模型的 AUC 都在 0.63–0.66 之間，略優於隨機（0.5），說明人口學特徵對多重用藥有一定解釋力但有限。

所有模型都嚴重偏向預測 Poly Drug（recall=89-93%），對 Single Drug 的 recall 極低（20-26%），這是 70:30 不平衡下的典型行為。



Feature Importance 顯示 `age` 和 `year` 是最重要的兩個特徵，說明多重用藥的傾向隨年齡和年份變化最大。

第 5 章｜非監督式學習

PCA 找的是資料中「變異量最大的方向」。17 個藥物特徵放在一起，有些可能高度相關（比如 fentanyl 和 fentanyl_analogue），PCA 會把它們合併成一個主成分，用更少的維度表達大部分資訊。

PCA 的結果顯示此資料集的特徵高度稀疏：需要 14 個主成分才能解釋 90% 的變異量（總共 17 個特徵），代表大多數藥物欄位各自獨立，沒有太多冗餘維度可壓縮。PC1 的 loadings 以 `drug_count`（0.674）、`fentanyl_analogue`（0.421）、`fentanyl_flag`（0.409）為主，代表第一主成分捕捉的是「fentanyl 類多重用藥強度」。

Apriori 關聯規則找到 46 條規則，最強的關聯（Lift=2.61）是

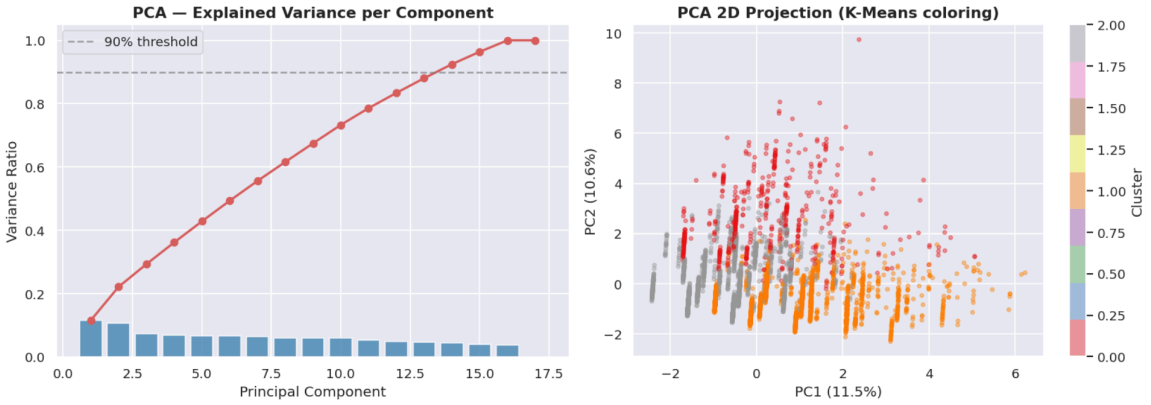
- `fentanyl_analogue → benzodiazepine + fentanyl`，以及
- `cocaine + fentanyl → fentanyl_analogue`

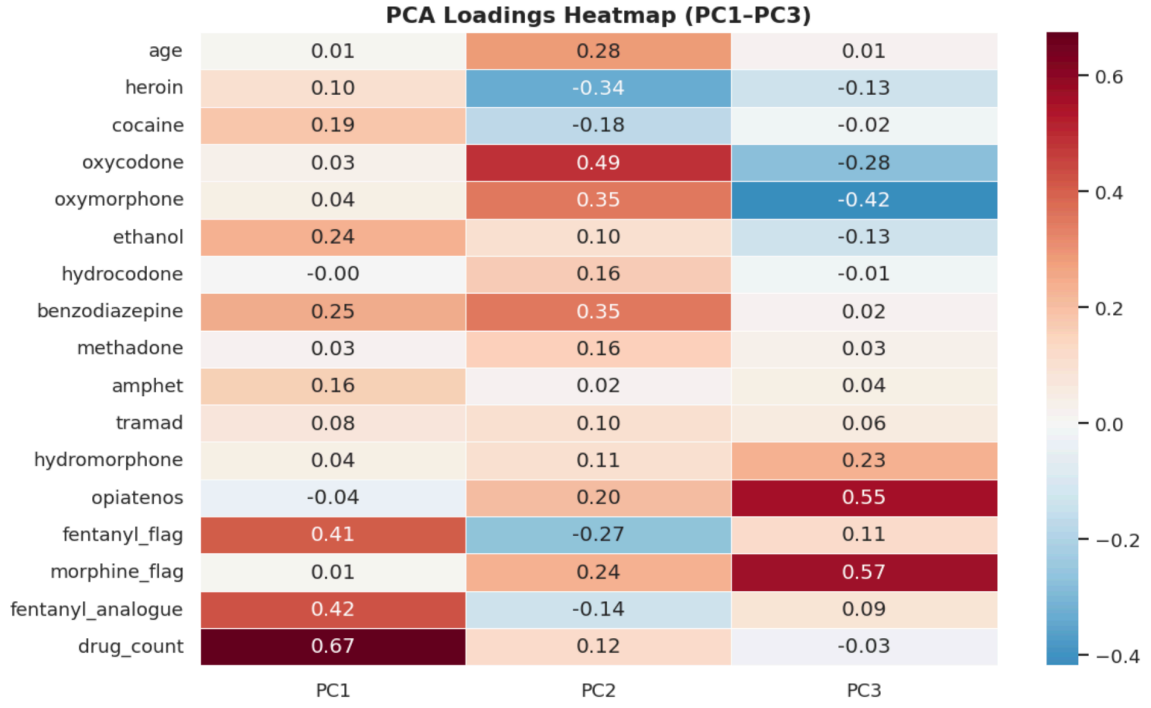
所有 confidence=100% 的規則都指向 fentanyl_flag（意思是凡涉及 fentanyl_analogue 的案例，100% 同時檢出 fentanyl 本體），這在藥理學上完全合理——fentanyl 類似物是 fentanyl 的衍生物。

Isolation Forest 偵測到 256 筆異常（5%），異常群的平均 drug_count=4.28（正常群=2.05），且幾乎所有藥物的涉及率都顯著高於正常群，代表異常案例就是「極端多重用藥」的個案。

```
特徵矩陣 Shape: (5105, 17)
特徵欄位 (17): ['age', 'heroin', 'cocaine', 'oxycodone', 'oxymorphone', 'ethanol', 'hydrocodone', 'benzodiazepin
e', 'methadone', 'amphet', 'tramad', 'hydromorphone', 'opiatenos', 'fentanyl_flag', 'morphine_flag', 'fentanyl_a
nalogue', 'drug_count']

達到 90% 解釋變異量所需主成分數: 14
```





PCA Loadings Heatmap（PC1–PC3）

- PC1（第一主成分，解釋 11.5%）

特徵	Loading	意義
drug_count	**+0.67**	最強，是整張圖的最大值
fentanyl_analogue	**+0.42**	芬太尼類似物
fentanyl_flag	**+0.41**	芬太尼本體
benzodiazepine	+0.25	安眠鎮定藥
ethanol	+0.24	酒精
cocaine	+0.19	古柯鹼

PC1 的正方向 = 用了越多種藥、越多芬太尼類藥物。
一個死者的 PC1 分數越高，代表他是越極端的多重用藥者，且幾乎必定涉及芬太尼。
注意 heroin 的 loading 只有 +0.10，說明海洛因本身跟「多重用藥強度」的關聯不強，用海洛因的人很多，但多數只用海洛因一種（Cluster 2 的 drug_count 平均只有 1.96）。

- PC2（第二主成分，解釋 10.6%）

特徵	Loading	意義
oxycodone	**+0.49**	最大正貢獻
oxymorphone	**+0.35**	正貢獻
benzodiazepine	+0.35	正貢獻
age	+0.28	年齡越大
heroin	**-0.34**	最大負貢獻
fentanyl_flag	-0.27	負貢獻
cocaine	-0.18	負貢獻

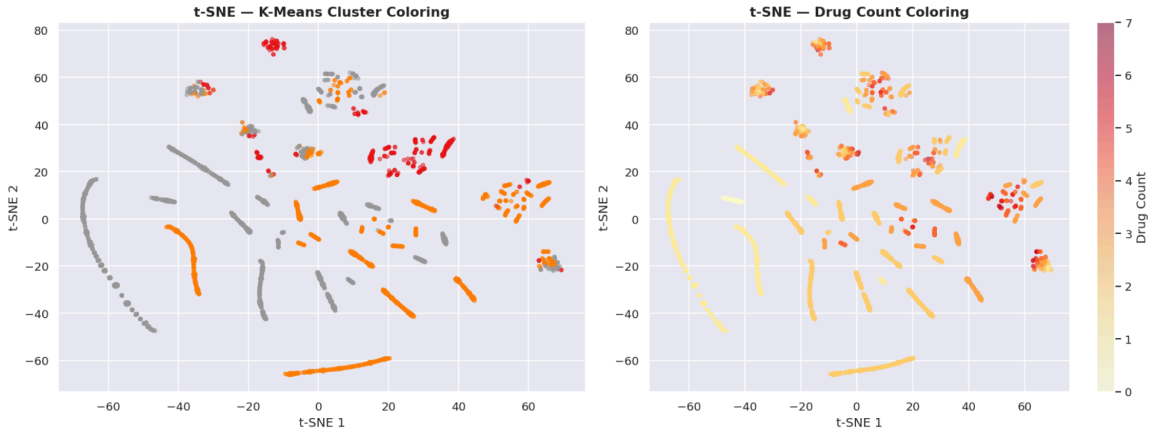
PC2 是一個「對立軸」：PC2 正方向 = 年紀較大、用處方止痛藥（oxycodone/oxymorphone）+ 安眠藥；PC2 負方向 = 用海洛因 + 芬太尼 + 古柯鹼（街頭毒品）。
這完美地量化了相關矩陣中 heroin ↔ oxycodone = -0.22 的發現，兩種截然不同的成癮路徑在 PC2 上被分到兩個反方向。

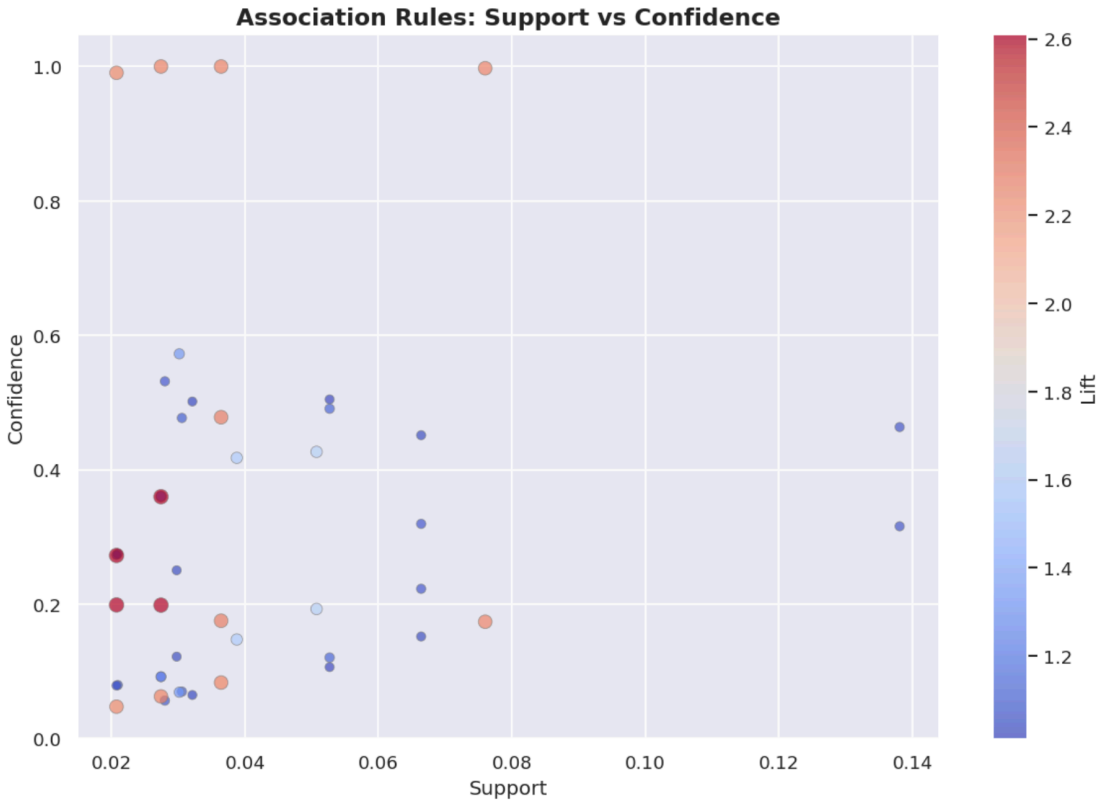
- PC3

特徵	Loading	意義
morphine_flag	**+0.57**	最大正貢獻
opiatenos	**+0.55**	非特定鴉片類
hydromorphone	+0.23	正貢獻
oxymorphone	**-0.42**	最大負貢獻
oxycodone	-0.28	負貢獻

PC3 的正方向 = 不明來源嗎啡 + 非特定鴉片類；負方向 = 具體的處方止痛藥（oxymorphone/oxycodone）。

這反映了法醫鑑定的不確定性，有些案例能明確鑑定到 oxycodone，有些只能說是「某種鴉片類」，PC3 把這兩種「鑑定清楚」vs「鑑定模糊」的案例分開了。

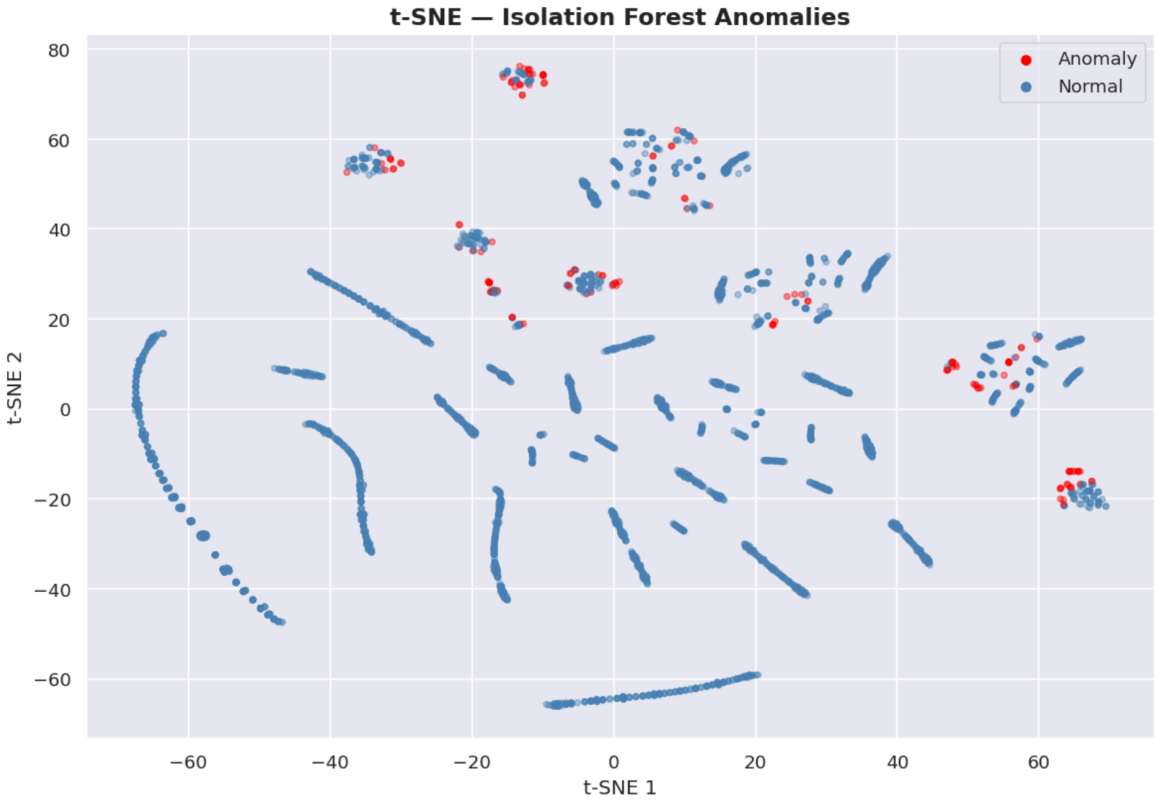




- 右上角（高 Support + 高 Confidence）： 沒有點。
- 左上角（低 Support + 高 Confidence）： 最有意思的區域。Support 只有 2–4%（罕見組合），但 Confidence 接近 1.0（幾乎確定）。最深紅的那幾個點（Lift ≈ 2.6）就在這裡，代表「雖然很少見，但一旦出現這種藥物組合，幾乎 100% 會同時出現另一種藥」。這些就是「芬太尼類似物 → 芬太尼本體」這類規則。
- 左下角（低 Support + 低 Confidence）： 點最密集的區域，大多是藍色（Lift 1.1–1.5）。Support 和 Confidence 都低，說明這些規則雖然存在，但關聯強度弱，實際意義有限。

- 右下角（高 Support + 低 Confidence）： 有兩個點在 Support ≈ 0.14（heroin 或 fentanyl 相關的規則），Confidence 只有 30–47%，代表即使是最常見的藥物組合，預測能力也不高——因為用海洛因的人太多了，他們的其他藥物組合非常多樣。

整體結論： 最有公衛意義的規則集中在「低 Support + 高 Confidence + 高 Lift」的左上角，全部指向芬太尼類似物相關的組合。



結論

這次專題分析了 Connecticut 州 2012–2018 年的意外藥物過量死亡資料，用了 EDA、分群、監督式與非監督式學習等方法來找出死亡案例的規律。

這次專題分析了 Connecticut 州 2012–2018 年的意外藥物過量死亡資料，用了 EDA、分群、監督式與非監督式學習等方法來找出死亡案例的規律。

資料顯示死亡人數從 2012 年一路飆升到 2017 年才稍微回落，而藥物種類上最大的變化是 fentanyl 後來居上，超越 heroin 成為主要致死藥物，反映出美國鴉片危機已經從傳統海洛因轉向合成鴉片類藥物。

死者以白人男性為主，中位年齡 42 歲，年齡分布出現雙峰，暗示年輕族群跟中年族群各有不同的用藥背景。

分群分析雖然沒找到很明顯的自然邊界，但還是可以歸納出幾種死亡型態，包括 fentanyl 類似物多重用藥、heroin 主導、美沙酮治療相關等，代表藥物過量死亡背後其實有很多條不同的路徑。

監督式學習的部分，三個模型的 AUC 都在 0.63–0.66 左右，表現差不多，說明光靠人口學和時間特徵還不夠用，加上資料本身類別不平衡，模型也容易偏向預測多重用藥。

這份資料最大的限制是只有驗屍紀錄，沒有死者生前的處方或用藥歷史，所以模型能學到的東西有限。如果未來能結合醫療或社經資料，分析結果應該會更有意義。