皮馬印地安人糖尿病風險預測：

建立早期診斷模型

組長：

112029056 林欣璇

組員：

112029002 江信亨

112029008 高翎毓

112029009 林 佳

112029011 張庭瑜

目錄

第一章　問題描述 **1**

第一節　資料來源與背景 1

第二節　資料集結構與分佈 1

第二章　問題假說與主題提出 3

第一節　研究主題 3

第二節　研究假設 3

第三章　資料前處理 4

第一節　高缺失率特徵移除 4

第二節　缺失值 4

第三節　離群值 4

第四節　KNN 補值 5

第四章　資料視覺化 8

第一節　熱力圖分析 8

第五章　規則建立 10

第一節　分析方法 10

第二節　實驗設計與模型驗證 10

第三節　模型評估與比較 13

第四節　綜合平均與最終測試結果 15

第六章　解決策略提出 19

第一節　閾值選取策略 19

第二節　最佳閾值 21

第七章　結論與討論 23

第八章　參考資料 24

附錄：程式碼說明 26

附錄：組員工作分配表、心得與照片 28

附錄：詳細的組員討論與訪談記錄表 32

附錄：上台報告時錄音筆所記下的問題答覆與說明 38

第一章　　問題描述

第一節　資料來源與背景

本研究使用之資料集最初源自美國國家糖尿病、消化和腎臟疾病研究所 (National Institute of Diabetes and Digestive and Kidney Diseases, NIDDK)。該資料集的建立目標是基於某些診斷測量值，來預測患者是否患有糖尿病。在樣本選擇上，該資料庫具有特定的人口統計限制：所有個案均為年滿 21 歲且具有皮馬印第安血統的女性。選擇此特定族群作為分析對象的原因在於，皮馬族印第安人已知為全球第二型糖尿病發病率最高的危險族群之一，具有高度的研究價值。

第二節　資料及結構與分佈

本資料集共包含 768 筆病患資料，每一筆資料由 8 個預測變數與 1 個目標變數組成。針對目標變數 Outcome 的分佈情形，本研究將病患區分為兩類：

- 陽性 (Positive, 1)：經診斷患有第二型糖尿病，共 268 位。
- 陰性 (Negative, 0)：未患有第二型糖尿病，共 500 位。

其中各欄位詳細定義與單位如下表一所示：

欄位名稱	中文說明	單位/備註	資料型態
Pregnancies	懷孕次數	次數	連續
Glucose	口服葡萄糖耐量試驗兩小時後靜脈血糖值	mg/dL	連續
BloodPressure	舒張壓	mm Hg	連續
SkinThickness	三頭肌的皮膚厚度	mm	連續
Insulin	兩小時後的血清胰島素	μU/ml	連續

	含量		
BMI	身體質量指數	kg / (m)²	連續
DiabetesPedigreeFunction	糖尿病血統函數	用於評估糖尿病遺傳可能性的評分函數	連續
Age	年齡	歲	類別
Outcome	是否患病	0 = 陰性 (無糖尿病) 1 = 陽性 (有糖尿病)	類別

表一

第二章　　問題假說與主題提出

第一節　研究主題

本研究旨在建構一個基於機器學習的糖尿病輔助診斷預測模型。透過分析皮馬印第安人女性的生理數據，探討各項臨床特徵與糖尿病發病之間的關聯性，並嘗試找出預測能力最佳的關鍵危險因子。

第二節　研究假說

在進行資料探勘模型訓練之前，本研究結合醫學領域知識與相關文獻回顧，針對皮馬印第安人糖尿病資料集提出以下兩點假說。這些假說將作為後續探索式資料分析的方向指引，並於模型評估階段進行驗證。

假說一：高風險因子之顯著關聯性

根據 Knowler 等人 (1981) 針對皮馬印第安人進行的長期追蹤研究，實證了肥胖與第二型糖尿病發病率之間存在強烈的正相關，且該族群的發病率遠高於其他族群。此外，美國糖尿病協會 (ADA, 2024) 的年度診療標準亦明確將 BMI 超標與空腹血糖異常列為篩檢糖尿病的首要指標。基於上述文獻，本研究推論血液中的葡萄糖濃度（Glucose）與身體質量指數（BMI）應與糖尿病的發病呈現高度正相關，即葡萄糖濃度越高、BMI 越高的族群，被診斷為陽性的機率將顯著增加。

假說二：機器學習模型之預測優勢

近期研究中，Zou 等人 (2018) 在比較多種演算法後，發現隨機森林等集成學習方法在糖尿病預測上通常能達到較單一模型更優異的表現。因此，本研究假說相較於單一指標的線性判斷，透過整合多項生理特徵的資料探勘演算法（如 Random Forest 或 XGBoost），應能有效捕捉變數間非線性的互動關係，從而達到更高的預測準確率。

第三章　　資料前處理

第一節　高缺失率特徵移除

經檢視資料缺失狀況，發現 Insulin 欄位的缺失比例極高（約 48.7% 的樣本值為 0）。根據 Hair 等人 (2010) 的建議，當單一變數的缺失值比例過高（通常超過 30%~50%）且無法藉由其他變數有效推估時，強行填補可能會引入過多雜訊與偏差。

此外，若從臨床病理學角度探討， 在第二型糖尿病發病初期，患者常伴隨胰島素阻抗現象，促使胰臟分泌大量胰島素以維持血糖恆定；然而隨著病程進展至後期，由於β細胞功能逐漸受損，胰島素濃度反而會顯著下降。此生理機制意味著胰島素數值與病程之間並非呈現單純的線性關係，而是具有高度的個體差異性。因此，為確保資料集的純淨度並維持預測模型的穩健性，本研究決定刪除 Insulin 欄位，以避免錯誤填補對模型造成誤導。

第二節　缺失值

針對 Glucose、BloodPressure 與 BMI 等欄位，雖然缺失比例相對較低，惟考量到生物醫學數據通常呈現偏態分佈且可能包含離群值，若直接使用平均數容易受到極端值影響。根據 Acuna 與 Rodriguez (2004) 的研究，對於非常態分佈的數據，使用中位數進行填補比平均數更具穩健性。因此，本研究採用中位數填補上述三個欄位之缺失值。

第三節　離群值

離群值是資料分布中明顯偏離其他觀測值的異常數值，若未處理將會影響資料分析的準確度。本研究採用四分位距法(IQR)進行離群值的計算與排除，四分位距 IQR=Q3-Q1，將小於 Q1−1.5×IQR 和大於 Q3+1.5×IQR 的值當作離群值，並將他們刪除，確保剩餘資料的集中性。

第四節　KNN 補值

在進行資料分析前，考量到原始資料中各變數具有不同的資料型態與數值分布範圍，若直接進行分析，極易導致模型偏重於數值較大的特徵，進而扭曲分析結果。特

別是對於依賴「距離計算」的模型（如 K-Nearest Neighbor, KNN），統一變數尺度至關重要。因此，本研究首先對所有連續型變數進行 0-1 正規化 (Normalization)，將原始數據依比例縮放至 $[0,1]$ 區間中，在不改變原始分佈情形的前提下，確保各變數在距離計算過程中具有一致的權重與尺度。正規化之具體操作步驟如圖一所示。

```
# 2. 建立 MinMaxScaler（將數值壓到[0, 1]）
scaler = MinMaxScaler()

# 3. 只對「數值欄位」做正規化
exclude_cols = ['Pregnancies', 'Outcome']
numeric_cols = df.select_dtypes(include=['int64', 'float64']).columns
numeric_cols = numeric_cols.difference(exclude_cols)

df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```

圖一

接下來針對 SkinThickness (皮膚皺褶厚度) 欄位，經檢視發現存在數值為 0 的不合理情況。為避免將其視為真實數值而誤導距離計算，本研究先將 SkinThickness 為 0 的樣本標記為缺失值，隨後採用 KNN 演算法進行插補。KNN 插補的核心概念是利用特徵空間中距離最近的 k 個樣本資訊來預測缺失值。本研究的實作流程（程式碼詳見圖二、圖三）包含以下步驟：

1. 數據正規化：確保距離計算的合理性。
2. 計算距離：針對含有缺失值的樣本，計算其與其他樣本間的歐式距離 (Euclidean Distance)。計算過程僅參考無缺失的特徵欄位。
3. 選擇 K 個最近鄰居：依據距離排序，選出最接近的樣本。
4. 填補缺失值：對於連續型特徵，採用鄰居的平均值或中位數填補；類別型特徵則採眾數填補。

```
# 3. 決定要補值的目標欄位
target = "SkinThickness"

# 用這些特徵（血糖、血壓、BMI、年齡、血緣功能）來抓KNN的距離
features = ["Glucose", "BloodPressure", "BMI", "Age", "DiabetesPedigreeFunction", target]

df_subset = df[features].copy()

df = df.replace({pd.NA: np.nan})
```

圖二

```
# 4. 建立 KNN 補值器
imputer = KNNImputer(n_neighbors=5)

# 5. 進行補值（這裡的 SkinThickness 是 [0, 1]正規化後的值）
df_imputed = pd.DataFrame(
    imputer.fit_transform(df_subset),
    columns=features
)
```

圖三

在本研究的具體設定中，目標填補項目為 SkinThickness，參考特徵包含 Glucose、BloodPressure、BMI、Age、DiabetesPedigreeFunction 以及 SkinThickness 本身。超參數設定方面，將鄰居數(k) 設定為 5，即由距離最近的 5 筆樣本進行估計；權重採用 Uniform Weighting (預設)，確保每一個近鄰樣本對補值結果具有相同影響力，避免結果過度受單一極近鄰樣本左右。

圖四（插補前）與 圖五（插補後）展示了資料插補前後的分佈對比。透過 KNN 算法，我們成功填補了缺失數值，同時保留了原始特徵的分佈趨勢，為後續模型訓練提供了完整且合理的數據基礎。

1	Pregnancies	Glucose	BloodPressure	SkinThickness	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	0.6709677419354838	0.5	0.5957446808510638	0.48427672955974843	0.24389160373167473	0.4833333333333334	1
3	1	0.264516129032258	0.40625	0.46808510638297873	0.26415094339622647	0.12127943136383827	0.16666666666666663	0
4	8	0.896774193548387	0.375		0.1603773584905661	0.2638827187916481	0.18333333333333335	1
5	1	0.2903225806451613	0.40625	0.3404255319148936	0.3113207547169812	0.03953798311861395	0.0	0
6	0	0.6000000000000001	0.0	0.5957446808510638	0.7830188679245284	0.9817858729453575	0.20000000000000007	1
7	5	0.4645161290322581	0.53125		0.23270440251572333	0.05464238116392714	0.15000000000000002	0
8	3	0.21935483870967737	0.15625	0.5319148936170213	0.4025157232704403	0.07552199022656596	0.08333333333333337	1
9	10	0.45806451612903226	0.5		0.5377358490566037	0.024877832074633496	0.13333333333333336	0
10	2	0.9870967741935484	0.46875	0.8085106382978723	0.3867924528301887	0.03553976010661927	0.5333333333333333	1
11	8	0.5225806451612902	0.875		0.44339622641509424	0.06841403820524211	0.55	1
12	4	0.42580645161290326	0.8125		0.610062893081761	0.05019991115059973	0.15000000000000002	0

圖四

	Pregnancies	Glucose	BloodPressure	SkinThickness	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	0.6709677419354838	0.5	0.5957446808510638	0.4842767295597484	0.2438916037316747	0.4833333333333334	1
3	1	0.264516129032258	0.40625	0.4680851063829787	0.2641509433962264	0.1212794313638382	0.1666666666666666	0
4	8	0.896774193548387	0.375	0.2638297872340425	0.1603773584905661	0.2638827187916481	0.1833333333333333	1
5	1	0.2903225806451613	0.40625	0.3404255319148936	0.3113207547169812	0.0395379831186139	0.0	0
6	0	0.6000000000000001	0.0	0.5957446808510638	0.7830188679245284	0.9817858729453576	0.2	1
7	5	0.4645161290322581	0.53125	0.3702127659574468	0.2327044025157233	0.0546423811639271	0.15	0
8	3	0.2193548387096773	0.15625	0.5319148936170213	0.4025157232704403	0.0755219902265659	0.0833333333333333	1
9	10	0.4580645161290322	0.5	0.5361702127659573	0.5377358490566037	0.0248778320746334	0.1333333333333333	0
10	2	0.9870967741935484	0.46875	0.8085106382978723	0.3867924528301887	0.0355397601066192	0.5333333333333333	1
11	8	0.5225806451612902	0.875	0.48085106382978715	0.4433962264150942	0.0684140382052421	0.55	1
12	4	0.4258064516129032	0.8125	0.5914893617021276	0.610062893081761	0.0501999111505997	0.15	0

圖五

第四章　　資料視覺化

第一節　熱力圖分析

在糖尿病確診群體（Outcome=1）中，可以觀察到「Skinthickness & BMI」、「Age & Pregnancies」以及「Age & BloodPressure」這三組成對關係中，都呈現較高的相關係數。

1. Skinthickness & BMI （三頭肌皮膚厚度及身體質量指數）

根據衛生福利部國民健康署之定義，18 歲以上成人體位區分為過輕（BMI ＜ 18.5）、正常（18.5 ≦ BMI ＜ 24.0）、過重（24.0 ≦ BMI ＜ 27.0）及肥胖（BMI ≧ 27.0）（衛生福利部，2018）。然而，BMI 的計算公式為體重（公斤）除以身高平方（公尺平方），其侷限性在於無法區分體內組織組成中之肌肉量與體脂肪占比。對於運動員或肌肉發達者，單一 BMI 指標可能導致肥胖程度之誤判與失真。

皮下脂肪厚度是評估個體脂肪堆積最直觀的臨床指標之一。搭配皮膚厚度測量以估算皮下脂肪比率，能有效彌補 BMI 的不足，提升肥胖辨識的準確性。本研究中，三頭肌皮膚厚度（SkinThickness）與 BMI 之相關係數 r 為 0.6，呈現顯著之中度正相關。值得注意的是，此結果與美國第三次國家健康與營養調查（NHANES III）人體測量報告數據呈高度一致。具體而言，根據 NHANES III 的大樣本分析顯示，成年男女三頭肌皮膚厚度與 BMI 之相關係數 r 通常落在 0.60 至 0.75 之間（平均值約為 0.65）。本研究測得之相關係數 r=0.69，不僅落在此合理區間內，更進一步反映了樣本族群在脂肪分佈上的一致性，證實了以皮膚厚度作為輔助診斷肥胖指標的臨床價值。

2. Age & Pregnancies（年齡與懷孕次數）

年齡與懷孕次數呈現中度正相關（r=0.44），從生理與社會學角度分析，隨著女性年齡增長，累積的懷孕次數自然隨之增加。然而，在糖尿病分析下，此相關性具有特殊的臨床意義。

根據 Li 等人（2016） 針對多項世代研究的統合分析（Meta-analysis）指出，女性的生產次數（Parity）與第二型糖尿病風險呈現顯著的劑量反應關係（Dose-response relationship），每增加一次生產，糖尿病風險便隨之提升。其生理機制在於，懷孕過程會導致母體經歷顯著的代謝壓力；胎盤分泌的荷爾蒙（如人類胎盤泌乳素）會生理性地增加胰島素阻抗，以確保胎兒營養供應。

Nicholson 等人（2006）也在文獻中提到，若婦女本身年齡較高或胰臟 β 細胞功能

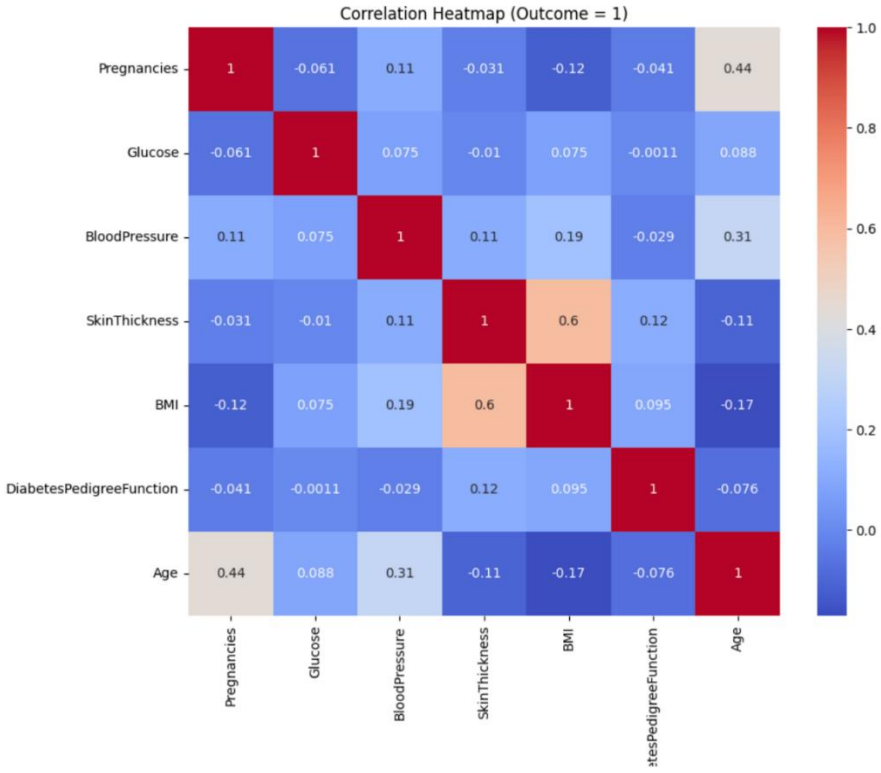
儲備不足，這種反覆且累積的「代謝壓力測試」，可能導致產後胰島素敏感度無法完全恢復。因此，本研究觀察到的相關性暗示了在確診族群中，「高齡且多產」是一個常見的特徵組合，這也與文獻中指出「多產次為糖尿病獨立風險因子」的結論相呼應。

3. Age & BloodPressure（年齡與血壓）

數據顯示，在糖尿病患者中，年齡與血壓呈現正相關（r=0.31）。這符合血管硬化與老化的生理進程。隨著年齡增長，血管壁彈性纖維逐漸流失，導致動脈硬化程度增加與收縮壓上升，這是老化的自然現象。

然而，糖尿病的存在加速了此一過程。Petrie 等人（2018） 的研究指出，高血糖狀態會引發氧化壓力（Oxidative Stress）與血管內皮細胞功能障礙（Endothelial Dysfunction），導致血管壁僵硬度比一般老化族群更為嚴重。此外，流行病學數據顯示，超過 75% 的第二型糖尿病患者同時合併有高血壓診斷。

因此，在 Outcome=1 的群體中觀察到年齡與血壓的連動上升，反映了「糖尿病」與「老化」雙重因素對心血管系統的加乘負擔。這提示了在糖尿病的照護模型中，針對高齡患者進行嚴格的血壓控制（Hypertension Control）是預防大血管與小血管併發症不可或缺的一環。



圖六

第五章　　解決策略提出

第一節　分析方法

在進行資料探勘模型的規則建立前，首先需依據 2.1 研究主題來釐清適用的分析技術。資料探勘中常見的規則建立方法主要包含「關聯分析 (Association Rule)」與「分類分析 (Classification)」兩類，其差異與本研究的選定依據如下：

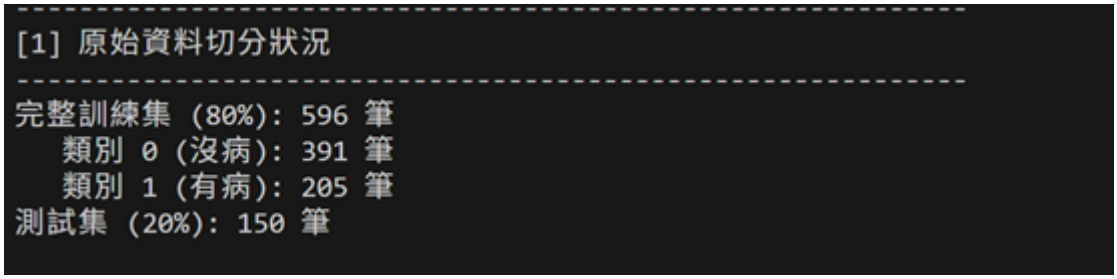
1. 關聯分析：主要用於發現資料項目之間同時出現的規律（例如：「購物籃分析」中，買了 A 的人通常也會買 B）。其產出形式為「若 X 發生，則 Y 發生的機率為 Z」，側重於描述變數間的共伴關係，而非對單一結果進行明確預測。
2. 分類分析：屬於監督式學習，旨在透過已知類別標籤的訓練資料，建立一個能將新資料準確分派至特定類別的函數模型。其產出形式為明確的預測結果（例如：患病 / 未患病），並能評估模型的準確率與變數重要性。

本研究選擇「分類分析」之原因則是呼應本研究 2.1 研究主題旨在建構「糖尿病輔助診斷預測模型」，我們的目標並非僅止於觀察生理特徵之間的共伴關係，而是需要一個能針對特定病患給出「是否罹患糖尿病（Outcome=0 or 1）」的明確決策。此外，根據 2.2 假說二，本研究預期透過集成式學習演算法來捕捉特徵間的非線性關係以提升預測優勢。相較於關聯規則，分類分析模型不僅能提供高準確度的預測，更能產出特徵重要性的排序，也能驗證假說中關於高風險因子的推論。

第二節　實驗設計與模型驗證

為了確保模型在類別不平衡資料上的泛化能力，並嚴格防止資料洩漏，本研究採用 5-Fold 分層交叉驗證進行評估。具體步驟如下：

1. **原始資料劃分：** 首先，將資料集依據 80:20 的比例進行劃分，切分狀況如圖。
    - **訓練集 (Training Set, 80%)**：用於模型訓練、特徵篩選與超參數調整。
    - **測試集 (Test Set, 20%)**：作為最終模型的獨立評估數據，在此階段完全隔離，不參與任何訓練或驗證過程，以確保最終測試結果的客觀性。



圖七

2. 分層五折交叉驗證：

針對上述劃分出的 80% 訓練集，本研究利用**分層 K 折交叉**驗證技術將其進一步劃分為 5 個子集（Folds）。此技術首先確保了每一折當中的類別比例（患病 vs. 健康）皆與原始訓練集保持一致。

隨後進行分層五折交叉驗證，其步驟為：在每一輪驗證循環中，依據 4:1 的比例，輪流選取其中 4 個子集（佔當前資料量約 80%）作為內部訓練集用於模型建構，而剩餘的 1 個子集（佔當前資料量約 20%）則作為內部驗證集進行效能評估。此過程共重複 5 次循環，確保訓練集中的每一筆資料都曾被當作驗證資料進行測試。

3. 參數最佳化：

在進行模型訓練之前，本研究採用網格搜索 (GridSearchCV)搭配五折交叉驗證 (5-Fold Cross Validation) 進行超參數最佳化。我們針對決策樹的數量 (n_estimators)、樹的最大深度 (max_depth) 以及葉節點最小樣本數 (min_samples_leaf) 進行了系統性的排列組合測試。

經過實驗比對，最終選定的最佳參數組合如下：

隨機森林(Random Forest) **(最佳化結果如圖八)**：

- 樹的最大深度 (max_depth): 10
- 決策樹的數量 (n_estimators): 200
- 葉節點最小樣本數 (min_samples_leaf): 4
- 最大特徵數 (max_features): sqrt
- 分裂最小樣本數 (min_samples_split): 2

針對上述參數設定，本研究特別說明隨機森林的變異性，這 200 棵決策樹 (n_estimators) 並非完全相同，其變異與多樣性主要來自兩個機制，這也是隨機森林能有效抗過擬合的關鍵，首先，在資料的隨機性上，每棵樹在訓練時，是透過「取後放回」的方式隨機抽取不同的訓練樣本，因此每棵樹所見的數據分佈略有不同；其次，關於特徵的隨機性，是透過設定 max_features: sqrt，限制每棵樹在進行節點分裂時，只能從總特徵數的開根號數量中隨機選取部分特徵進行評估。這強迫每棵樹從不同特徵角度切入分析，確保了模型整體的多樣性。

XGBoost **(最佳化結果如圖九)**：

- 樹的最大深度 (max_depth): 2
- 決策樹的數量 (n_estimators): 150
- 隨機抽樣比例 (subsample): 0.7
- 節點分裂的最小損失下降 (gamma): 2 (通常從 0 開始往上調，如果模型過擬合可以再往上調；但如果 gamma 調太大，模型可能會欠擬合)
- 學習率 (learning rate): 0.05

```
最終選定最佳參數: {'rf__max_depth': 10, 'rf__max_features': 'sqrt', 'rf__min_samples_leaf': 4, 'rf__min_samples_split': 2, 'rf__n_est
imators': 200}
最佳 CV Score (roc_auc): 0.8389
```

圖八

```
===== 正在進行 Grid Search 優化參數... =====
最佳參數組合: {'gamma': 2, 'learning_rate': 0.05, 'max_depth': 2, 'n_estimators': 150, 'subsample': 0.7}
```

圖九

4. 動態平衡處理 ：

為了解決樣本不平衡問題，本研究使用 Borderline-SMOTE，在此步驟人工合成樣本僅針對當下的訓練集進行生成，將少數類別（患病）擴增至與多數類別（健康）的數量一致。另外，考量醫療數據中，患病與健康樣本在特徵空間上常存在重疊與模糊地帶。傳統 SMOTE 易造成雜訊生成或過度泛化；反之，Borderline-SMOTE 能自動識別並專注於強化決策邊界上的難分類樣本 。此策略能有效釐清類別界線，使模型更精準地學習區分處於灰色地帶的潛在病患，從而提升預測效能，如圖十和圖十一。關鍵在於，在此過程中驗證集

（Validation Set）完全不參與合成過程，如圖十二和圖十三，驗證集始終保持原始真實分佈，以確保評估結果的公正性。

隨機森林Train	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SMOTE前	0=312, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164
SMOTE後	0=312, 1=312	0=313, 1=313	0=313, 1=313	0=313, 1=313	0=313, 1=313

圖十

XG Boost Train	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SMOTE前	0=312, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164
SMOTE後	0=312, 1=312	0=313, 1=313	0=313, 1=313	0=313, 1=313	0=313, 1=313

圖十一

隨機森林Val	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
set	0=79, 1=41	0=78, 1=41	0=78, 1=41	0=78, 1=41	0=78, 1=41

圖十二

XG Boost Val	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
set	0=79, 1=41	0=78, 1=41	0=78, 1=41	0=78, 1=41	0=78, 1=41

圖十三

5. 模型訓練與評估：使用經平衡處理後的訓練集訓練模型，隨後使用未經處理的驗證集進行預測。評估指標包含 AUC、Accuracy 與 F1-Score，以多面向衡量模型效能。

6. 最終效能指標：計算各項指標在 5 折訓練集和驗證集中的平均值與標準差，作為該模型最終的效能評估依據，以降低單次切分帶來的隨機誤差。

第三節　模型評估與比較

運用前述篩選出的最佳參數組合，本研究分別對 XGBoost 與 Random Forest 進行了完整的模型訓練與評估。為了全面衡量模型在類別不平衡資料下的預測能力，本研究採用以下三項指標進行綜合比較：

1. AUC：評估模型區分正負樣本（患病 vs. 健康）的整體能力，數值越接近 1 代表區別力越強。
2. Accuracy ：衡量整體預測正確的比例。
3. F1-Score：考量 Precision（精確率）與 Recall（召回率）的調和平均數，特別適用於觀察模型對少數類別（正樣本）的捕捉成效。

在計算平均效能之前，本研究首先詳細檢視了模型在五折交叉驗證中每一折（Fold）的具體表現，以確保模型並非僅在特定資料劃分下表現優異，而是具有穩定的預測能力。

首先觀察隨機森林的表現，**圖十四**為隨機森林在訓練集（Train Set）的五折數據，可以看出模型在學習階段表現極為強勢，五次切分的 AUC 皆穩定維持在 0.98 以上，F1-Score 也都在 0.85 至 0.90 之間，顯示模型對訓練資料有極佳的擬合度。

然而，當我們檢視**圖十五**的驗證集（Validation Set）結果時，可以觀察到明顯的效能落差。雖然 Accuracy 維持在 75% 左右，但在關鍵指標 F1-Score 上，五折的數據出現了波動（介於 0.65 到 0.72 之間）。此外，觀察圖十五下方的混淆矩陣（Confusion Matrix），隨機森林在判定正樣本（Type 1，有病）時，雖然漏判（FN）較少，但將健康誤判為有病（FP）的情況相對較多（例如 Fold 1 的 FP 高達 17）。

隨機森林Train	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SMOTE前	0=312, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164
SMOTE後	0=312, 1=312	0=313, 1=313	0=313, 1=313	0=313, 1=313	0=313, 1=313
指標	0.9097	0.9308	0.8910	0.9245	0.9015
Accuracy	0.9874	0.9878	0.9804	0.9837	0.9779
AUC					
F1	0.8802	0.9054	0.8571	0.8977	0.8691

圖十四

隨機森林Val	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
set	0=79, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41	
指標	0.7917		0.7395		0.7479		0.7479		0.7563	
Accuracy	0.8524		0.8018		0.8215		0.8008		0.8399	
AUC										
F1	0.7253		0.6931		0.6591		0.6809		0.6882	
混淆矩陣	33	17	35	25	29	18	32	21	32	20
TP FP										
FN TN	8	62	6	53	12	60	9	57	9	58

圖十五

接著觀察 XGBoost 的表現，**圖十六**顯示 XGBoost 在訓練集的 AUC 約落在 0.93 上下，雖然在數值上略低於隨機森林的 0.98，但這反而是一個健康的訊號，暗示模型沒有過度死記訓練資料。

這一點在**圖十七**的驗證集表現中得到了證實，XGBoost 在驗證集的五折表現中，AUC 穩定維持在 0.81 到 0.87 的區間，且 F1-Score 的表現優於隨機森林。更重要的是，觀察圖十七的混淆矩陣，XGBoost 在各個 Fold 中的預測分佈更為均衡，能在保持較高 True Positive (TP) 的同時，有效控制錯誤率。這證明了 XGBoost 採用的序列式提升策略在處理此類別不平衡資料時，具有更佳的穩定性與泛化潛力。

XG Boost Train	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SMOTE前	0=312, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164
SMOTE後	0=312, 1=312	0=313, 1=313	0=313, 1=313	0=313, 1=313	0=313, 1=313
指標	0.8638	0.8770	0.8594	0.8658	0.8578
Accuracy	0.9340	0.9454	0.9287	0.9355	0.9290
AUC					
F1	0.8686	0.8821	0.8654	0.8679	0.8620

圖十六

XG Boost Val	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
set	0=79, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41	
指標	0.8333		0.7311		0.7899		0.7479		0.7731	
Accuracy	0.8740		0.8161		0.8405		0.8124		0.8674	
AUC										
F1	0.7619		0.6667		0.7126		0.6739		0.6966	
混淆矩陣	68	11	55	23	63	15	58	20	61	17
TP FP										
FN TN	9	32	9	32	10	31	10	31	10	31

圖十七

第四節　綜合平均與最終測試結果

綜合上述五折的詳細數據，我們將其平均後整理如**表二**與**表三**。

如**表二**所示，兩種模型在訓練集上皆展現了極佳的學習能力，其中 Random Forest 的平均 AUC 高達 0.9834，略高於 XGBoost 的 0.9345。然而，在更能反映模型穩定性的驗證集中（**表三**），兩者的指標均有所下降，顯示出小樣本醫療數據常見的過擬合挑戰。在此階段，XGBoost 在 AUC、Accuracy 與 F1-Score 三個指標上的平均表現皆略優於 Random Forest，且標準差（Standard Deviation）控制在合理範圍，呼應了前述分折檢視中觀察到的穩定性。

模型演算法	訓練集 AUC	訓練集 Accuracy	訓練集 F1-Score
XGBoost	0.9345 ± 0.0061	0.8648 ± 0.0068	0.8692 ± 0.0068
Random Forest	0.9834 ± 0.0039	0.9115 ± 0.0146	0.8819 ± 0.0178

表二

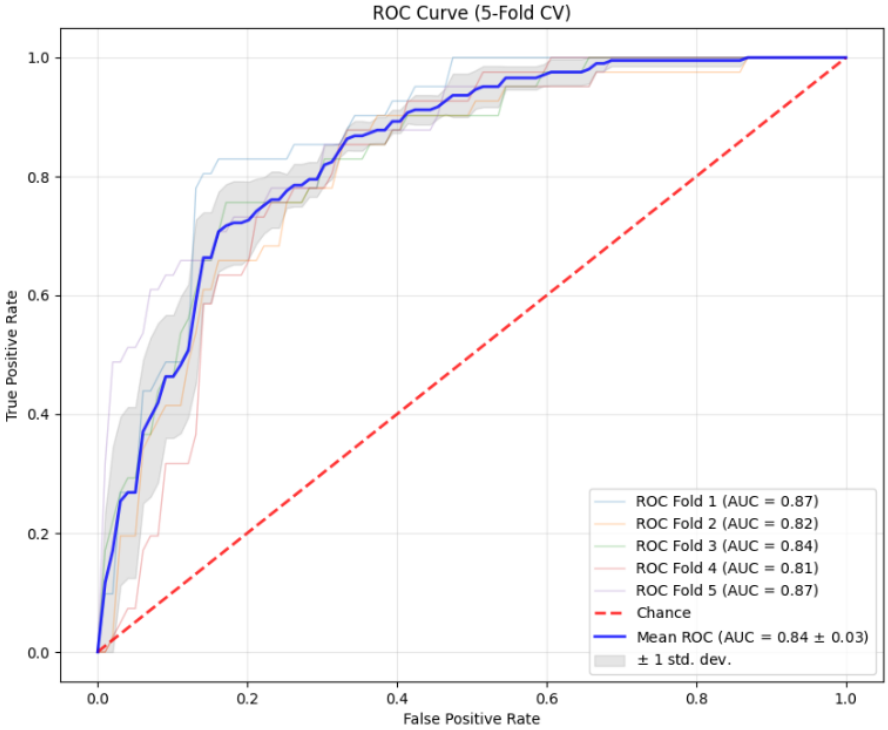
模型演算法	驗證集 AUC	驗證集 Accuracy	驗證集 F1-Score
XGBoost	0.8421 ± 0.0254	0.7751 ± 0.0355	0.7024 ± 0.0340
Random Forest	0.8233+-0.0205	0.7567+-0.0183	0.6893+-0.0214

表三

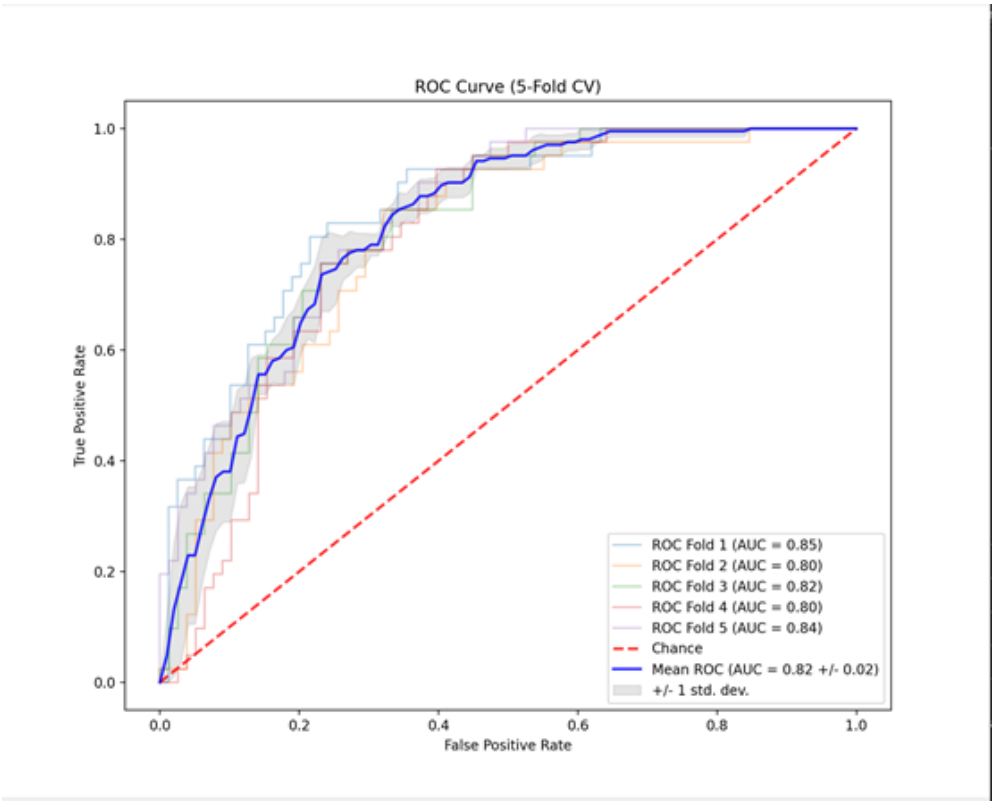
為了確認模型的泛化能力，我們使用完全未參與訓練的 20% 獨立測試集進行評估，結果如表四所示。值得注意的是，XGBoost 展現了更優異的泛化能力。其測試集 AUC 達到 0.8399，優於 Random Forest 的 0.8106；且在關鍵的 F1-Score (0.7288) 上也取得了最佳成績。這顯示 XGBoost 在面對未知數據時，對於潛在糖尿病患的預測更為精準且穩健。

	測試集 AUC	測試集 Accuracy	測試集 F1-Score
XGBoost	0.8399	0.7867	0.7288
Random Forest	0.8106	0.7467	0.6935

表四


ROC Curve (5-Fold CV)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Fold 1 (AUC = 0.87)
ROC Fold 2 (AUC = 0.82)
ROC Fold 3 (AUC = 0.84)
ROC Fold 4 (AUC = 0.81)
ROC Fold 5 (AUC = 0.87)
Chance
Mean ROC (AUC = 0.84 ± 0.03)
± 1 std. dev.

圖十四


ROC Curve (5-Fold CV)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Fold 1 (AUC = 0.85)
ROC Fold 2 (AUC = 0.80)
ROC Fold 3 (AUC = 0.82)
ROC Fold 4 (AUC = 0.80)
ROC Fold 5 (AUC = 0.84)
Chance
Mean ROC (AUC = 0.82 +/- 0.02)
+/- 1 std. dev.

圖十五

最後，對比兩張 ROC 曲線圖，可以看出 XGBoost 的整體曲線（圖十四藍色實線）相較於 Random Forest 的曲線（圖十五藍色實線），更往左上角（高真陽性率、低偽陽性率區域）貼近。這顯示在相同的偽陽性率 (False Positive Rate) 下，XGBoost 通常能達到更高的真陽性率 (True Positive Rate)，具備更佳的分類能力。

綜合上述實驗結果與指標分析，本研究最終選定 **XGBoost** 作為核心預測模型，主要基於以下兩大理由：

1. **整體預測效能較佳**：在嚴謹的 5-Fold 交叉驗證比較中，XGBoost 的平均 AUC (0.842) 顯著高於 Random Forest (0.823)，顯示其在特徵學習與分類判斷上更為精準。
2. **穩健的泛化能力**：XGBoost 在驗證集與測試集之間的表現極為穩定 (AUC 0.842 vs. 0.840)，證明該模型具有高度的可靠性，適合應用於實際的預測情境。

綜上所述，XGBoost 在各項關鍵指標上均優於 Random Forest，因此本研究決議採用 XGBoost 進行後續的最佳閾值分析與預測應用。

第六章　　解決策略提出

第一節　閾值選取策略

1. 前言與閾值選取策略

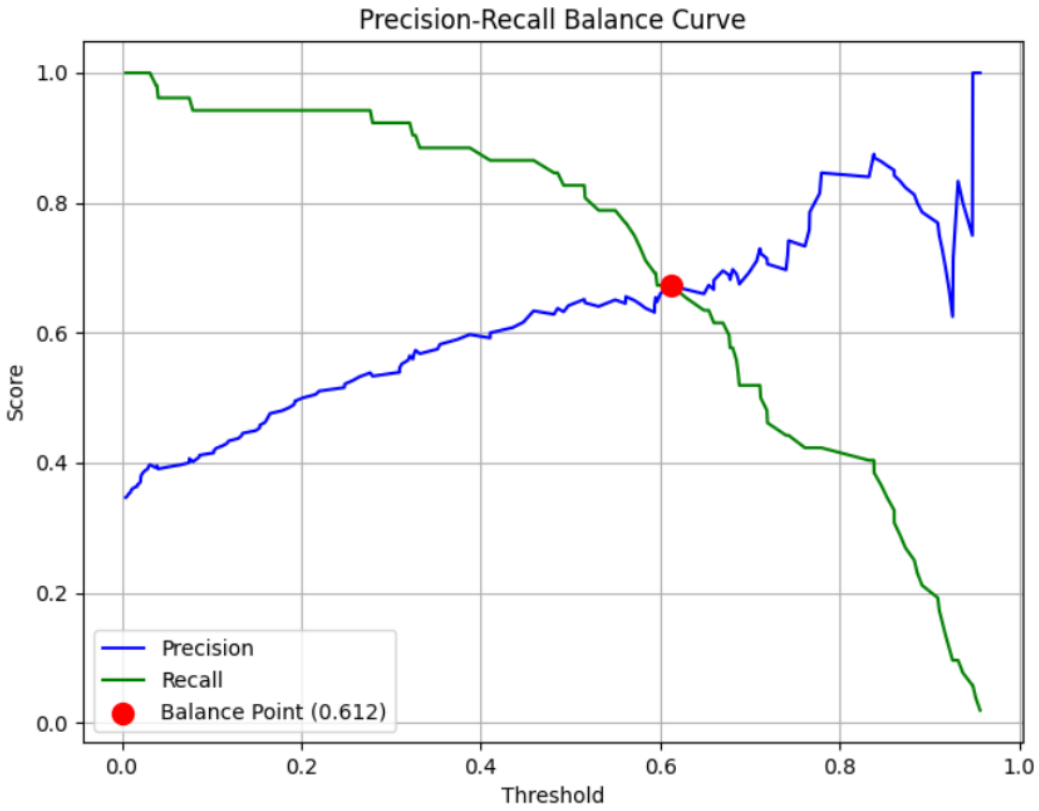
當模型的測試集 AUC 已達到 0.8 以上，且 Accuracy 也趨於穩定，本研究轉而將目標轉向 outcome 機率值的閾值調整，以完善模型的診斷能力。本研究首先使用預設閾值(Threshold=0.5) 完成 XGBoost 模型的訓練與初步評估。其原始輸出為對數勝算(Log-Odds)，需經由 Sigmoid 函數轉換為介於 0 與 1 之間的機率值。根據標準決策理論，分類器的預設判斷規則遵循 Sgn{(sigmoid(x)) - 0.5}，即當預測機率 P(y=1|x) > 0.5 時判定為陽性，反之則為陰性。

然而，根據 He, H., & Garcia, E. A. (2009).的研究，在醫療診斷或疾病預測的應用中，資料集往往伴隨著類別不平衡的問題，且偽陽性 (FP) 與偽陰性 (FN) 所帶來的代價並不相等 。因此，根據 Siers, M. J., & Islam, M. Z. (2015)指出閾值調整在醫療分類的重要性，我們認為單純依賴預設的 0.5 切點未必符合臨床需求。為了優化診斷效能，本節採用「Precision-Recall 平衡策略」找出數學上的最佳閾值，並將此新閾值重新應用於測試集 (Test Set) 的預測機率——將機率高於最佳閾值的樣本重新歸類為 1，其餘為 0——藉此觀察在更嚴謹的判斷標準下，模型對於誤判控制的表現。

2. 閾值分析結果

如圖十二所示，我們繪製了精確率 (Precision) 與召回率 (Recall) 隨閾值變化的曲線，以視覺化方式呈現模型在不同決策標準下的效能消長；**橫軸**代表分類閾值 (Threshold)，範圍由 0 至 1；**藍色曲線**代表 Precision。隨著閾值提高，模型判定陽性的標準變嚴格，誤判 (FP) 減少，因此 Precision 呈現上升趨勢；**綠色曲線**代表 Recall。隨著閾值提高，部分真實陽性樣本被排除，導致 Recall 呈現下降趨勢。

本研究依據「Precision-Recall 平衡策略」，尋找兩條曲線的交點（即圖中紅點所示）。分析結果顯示，當閾值設定為 **0.612** 時，Precision 與 Recall 達到平衡狀態，此時兩者數值皆約為 **0.673**。此交叉點代表模型在「捕捉病患能力 (Recall)」與「診斷準確度 (Precision)」之間取得了數學上的最佳折衷。因此，我們選定 **0.612** 作為調整後的最佳閾值。



圖十二

3. 效能比較：預設閾值 vs. 平衡閾值

為了評估調整閾值後的效益，我們將測試集 (Test Set) 在預設閾值 (Th=0.5) 與平衡閾值 (Th=0.612) 下的表現進行了比較，詳細數據如表七 所示。

評估指標	預設閾值 (0.5)	平衡閾值 (0.612)	變化趨勢
AUC	0.8399	0.8399	不變 (與閾值無關)
Accuracy	0.7867	0.7733	下降 1.34%
F1-Score	0.7288	0.6731	下降
Precision	0.6515	0.6731	上升

Recall	0.8269	0.6731	下降

表七

接下來根據混淆矩陣深入分析，進一步檢視混淆矩陣的變化，可觀察到閾值移動對錯誤類型分布的直接影響：

- 偽陽性 (False Positive, FP)：由 23 例下降至 17 例，其臨床意義表示有 6 位原本被誤判為糖尿病的健康受測者，在調整閾值後被正確排除。這直接降低了醫療資源的無效投入以及受測者不必要的心理負擔。
- 偽陰性 (False Negative, FN)：由 9 例上升至 17 例，其臨床意義表示隨著 FP 的改善，漏判個案同時增加。

透過將閾值從 0.5 調整至 0.612，模型成功將 Precision 提升至與 Recall 平衡的水準，並顯著壓低了誤判數 (FP)。然而，數學上的等值點 (Precision=Recall) 雖然客觀，卻未必完全符合特定醫療場域對於「風險趨避」的偏好。因此，下一節將以此等值點為基礎，進一步進行「手動閾值微調」，以探索在臨床實務上更具操作價值的最佳切點。

第二節　最佳閾值

依據 Provost 與 Fawcett (2001) 提出的成本敏感學習 (Cost-Sensitive Learning) 理論，最佳分類閾值的選取不應僅依賴統計指標，而必須考量在特定應用情境下，誤判（FP）與漏判（FN）所造成的後果嚴重性 。本研究將所開發之模型定位為「早期輔助診斷支援系統 」， 一般篩檢工具首重高敏感度以防漏判，但在早期診斷的情境中，模型的首要任務是提供具備高信賴度的陽性預警，以協助醫師進行診斷決策。因此，我們認為「降低偽陽性 (False Positive, FP)」具有高度優先性，以避免健康受測者因系統誤報而產生不必要的心理恐慌，並減少後續無效醫療資源的投入。

然而，追求低誤判的同時仍需維持模型最基本的預測效力。我們參考 Power 等人 (2013) 提出的診斷測試法則，認為一個具備實用價值的模型，如果 Sensitivity 掉到 0.5，那相當於是盲目亂猜。因此，本研究主要想測試在將閾值繼續調高的同時，是否能達到誤判人數(FP)降低，並且仍能維持 Recall 的判斷能力，確保模型捕捉病患的能力高於隨機機率（Random Chance, 0.50） 。

然而，追求低誤判的同時仍需兼顧基本的篩檢能力。隨著閾值提高，雖然 FP 得以壓低，但偽陰性（FN）的增加速率終將超越 FP 的下降效益。根據圖十三的閾值分析顯示：

1. 安全區間 (0.6120 -> 0.6820)：

   當閾值提升至 0.6820 時，雖然 FN 增 5 例，但成功使 FP 減少 4 例。更關鍵的是，此時召回率 (Recall) 維持在 0.5577，仍大於隨機亂猜的 0.5。

2. 危險區間 (0.7020 以上)：

   若試圖將閾值進一步提升至 0.7020 或更高，雖然能再減少 2 例 FP，但 Recall 將急劇下降至 0.5192 甚至跌破 0.5。這顯示此時犧牲 recall 以換取 precision 的代價過高，已失去篩檢意義。

```
Threshold  | FP (誤判)  | FN (漏判)  | Precision  | Recall     | F1
--------------------------------------------------------------------------
0.6120     | 17         | 18         | 0.6667     | 0.6538     | 0.6602
0.6220     | 17         | 18         | 0.6667     | 0.6538     | 0.6602
0.6320     | 17         | 19         | 0.6600     | 0.6346     | 0.6471
0.6420     | 17         | 19         | 0.6600     | 0.6346     | 0.6471
0.6520     | 16         | 19         | 0.6735     | 0.6346     | 0.6535
0.6620     | 14         | 20         | 0.6957     | 0.6154     | 0.6531
0.6720     | 14         | 21         | 0.6889     | 0.5962     | 0.6392
0.6820     | 13         | 23         | 0.6905     | 0.5577     | 0.6170
0.6920     | 12         | 25         | 0.6923     | 0.5192     | 0.5934
0.7020     | 11         | 25         | 0.7105     | 0.5192     | 0.6000
0.7120     | 10         | 27         | 0.7143     | 0.4808     | 0.5747
   *** 警告: Recall 已低於 0.5 (0.4808) ***
0.7220     | 10         | 29         | 0.6970     | 0.4423     | 0.5412
   *** 警告: Recall 已低於 0.5 (0.4423) ***
0.7320     | 10         | 29         | 0.6970     | 0.4423     | 0.5412
   *** 警告: Recall 已低於 0.5 (0.4423) ***
0.7420     | 8          | 29         | 0.7419     | 0.4423     | 0.5542
   *** 警告: Recall 已低於 0.5 (0.4423) ***
0.7520     | 8          | 30         | 0.7333     | 0.4231     | 0.5366
   *** 警告: Recall 已低於 0.5 (0.4231) ***
0.7620     | 7          | 30         | 0.7586     | 0.4231     | 0.5432
   *** 警告: Recall 已低於 0.5 (0.4231) ***
0.7720     | 5          | 30         | 0.8148     | 0.4231     | 0.5570
   *** 警告: Recall 已低於 0.5 (0.4231) ***
0.7820     | 4          | 31         | 0.8400     | 0.4038     | 0.5455
   *** 警告: Recall 已低於 0.5 (0.4038) ***
0.7920     | 4          | 31         | 0.8400     | 0.4038     | 0.5455
   *** 警告: Recall 已低於 0.5 (0.4038) ***
PS D:\DM project>
```

圖十三

因此基於「輔助診斷定位」與「誤判極小化」的雙重考量，本研究最終不再上調閾值，並將其保持在 0.6012，以維持模型的判定水準。

第七章　　結論與討論

本研究成功建構了**具備良好預測效能**的糖尿病輔助診斷模型，經由嚴謹的五折交叉驗證評估，結果顯示 XGBoost 演算法在各項指標表現上皆優於隨機森林（Random Forest）；特別是在測試集上，XGBoost 的 AUC 值達到 0.8399，證實了模型具備優異的預測準確度與泛化能力。

在臨床應用策略上，為了優化篩檢效益，本研究採用 Precision-Recall 平衡策略進行分析，最終將模型判定閾值設定為 0.612。在此設定下，系統成功在「避免遺漏病患」與「降低對健康民眾誤判」的雙重目標中取得最佳數學平衡點，大幅提升了輔助診斷的參考價值。此外，特徵重要性分析指出 Glucose（血糖）、Age（年齡）及 BMI 為影響預測的最關鍵因子，此數據結果不僅與 ADA 診療標準高度一致，亦符合 Knowler (1981) 的文獻發現，證實本模型所學習到的邏輯完全符合真實病理機制，兼具數據精準度與醫學解釋力。

第八章　　參考資料

1. American Diabetes Association. (2024). Standards of Medical Care in Diabetes—2024. Diabetes Care, 47(Supplement_1).
2. Knowler, W. C., Pettitt, D. J., Savage, P. J., & Bennett, P. H. (1981). Diabetes incidence in Pima Indians: contributions of obesity and parental diabetes. American Journal of Epidemiology, 113(2), 144-156.
3. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515.
4. python 資料處理-缺失值處理基礎操作語法彙整
5. 資料正規化(Normalization)與標準化(Standardization)
6. Durnin, J. V., & Womersley, J. V. G. A. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *British journal of nutrition*, *32*(1), 77-97.doi: 10.1079/bjn19740060. PMID: 4843734.
7. http://api.lib.ntnu.edu.tw:8080/server/api/core/bitstreams/ee782f0f-6ed9-4742-8061-c558afe9353b/content
8. https://medium.com/@whchang022/%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-a939adfa96de
9. 機器學習任務：分類！迴歸！分群！
10. https://www.pcschool.com.tw/blog/it/supervised-learning
11. https://hackmd.io/@CynthiaChuang/Common-Evaluation-MetricAccuracy-Precision-Recall-F1-ROCAUC-and-PRAUC
12. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.
13. Siers, M. J., & Islam, M. Z. (2015). Structural bias in binary classification on imbalanced datasets: Threshold adjustment vs. resampling. In Australasian Joint Conference on Artificial Intelligence
14. Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3), 203-231.
15. Power, M., Fell, G., & Wright, M. (2013). Principles for high-quality, high-value testing. *Evidence-Based Medicine*, 18(1), 5-10.
16. Li, W., Ruan, W., Peng, Y., & Wang, D. (2016). Parity and risk of type 2 diabetes: a dose-response meta-analysis of cohort studies. *Scientific Reports*, 6, 22153.
17. Nicholson, W. K., Asao, K., Brancati, F., Coresh, J., & Powe, N. R. (2006). Parity and the metabolic syndrome in US women. *Diabetes Care*, 29(6), 1249-1254.

18. Petrie, J. R., Guzik, T. J., & Touyz, R. M. (2018). Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Canadian Journal of Cardiology*, 34(5), 575-584.
19. Long, A. N., & Dagogo-Jack, S. (2011). Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. *The Journal of Clinical Hypertension*, 13(4), 244-251.
20. Durnin, J. V., & Womersley, J. V. G. A. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *British journal of nutrition*, *32*(1), 77-97.doi: 10.1079/bjn19740060. PMID: 4843734.

附錄: 程式碼說明

本研究使用 Python 語言結合 scikit-learn 與 XGBoost 套件進行實作。為確保實驗結果的穩健性與可重現性（Reproducibility），程式碼設計涵蓋了資料分層切分、防止資料洩漏的動態過採樣（Dynamic SMOTE）、五折交叉驗證以及閾值優化策略。以下說明關鍵程式邏輯：

1. 資料集切分與分層抽樣

為避免因資料不平衡導致訓練集與測試集分佈不均，本研究採用分層抽樣（Stratified Sampling）將資料劃分為 80% 訓練集與 20% 測試集，確保兩者間的類別比例（Outcome 0 與 1）一致。

程式碼片段：分層切分

```
X_train_all, X_test, y_train_all, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)
```

2. 防止資料洩漏的五折交叉驗證

本研究最關鍵的實作細節在於 SMOTE 的應用時機。為了避免「資料洩漏（Data Leakage）」，我們並未在整份資料上先進行 SMOTE，而是將 SMOTE 放入交叉驗證的每一折（Fold）中，在每一折訓練時，僅對當下的「訓練子集」進行過採樣生成合成資料，而「驗證子集」保持原始真實分佈，其操作是為了確保模型是在沒看過驗證資料的情況下進行評估，反映真實的泛化能力。

```
for fold, (train_idx, val_idx) in enumerate(skf.split(X_train_all, y_train_all), 1):
    print(f"\n>>>>> Running Fold {fold} / 5 <<<<<")

    # 切分 Fold
    X_tr, X_val = X_train_all.iloc[train_idx], X_train_all.iloc[val_idx]
    y_tr, y_val = y_train_all.iloc[train_idx], y_train_all.iloc[val_idx]

    # SMOTE
    smote = SMOTE(random_state=42)
    X_tr_res, y_tr_res = smote.fit_resample(X_tr, y_tr)

    # 訓練模型
    model = XGBClassifier(**model_params)
    model.fit(X_tr_res, y_tr_res)
```

3. 最佳閾值優化

有別於傳統預設的 0.5 門檻值，本程式透過 precision_recall_curve 計算不同閾值下的 Precision 與 Recall，並自動尋找兩者最接近的交點，由程式碼可以看到，尋找 abs(Precision - Recall) 最小處，並讓程式自動輸出的最佳閾值（如前述結論中的 0.612），即為此邏輯之計算結果。

程式碼片段：尋找最佳平衡閾值

```
precision, recall, thresholds = precision_recall_curve(y_test, y_test_prob)
diff = np.abs(precision[:-1] - recall[:-1])
best_idx = np.argmin(diff)
best_threshold = thresholds[best_idx]
```

附錄:組員工作分配表、心得與照片

名字	分工
112029002 江信亨	特徵重要性分析
112029008 高翎毓	PPT 製作、數據整理
112029009 林 佳	模型選擇、閾值調整與模型訓練
112029011 張庭瑜	資料前處理、資料視覺化、閾值調整與模型訓練、文書撰寫
112029056 林欣璇	資料前處理、模型選擇、閾值調整與模型訓練、文書撰寫

112029002 江信亨

這學期修資料探勘這門課，對我來說並不算輕鬆，很多內容都覺得有點吃力，也沒有真正學會全部的技術。一開始我對資料分析的理解比較偏向統計，會覺得就是用一些方法算數字，再看結果怎麼解釋。但做了糖尿病風險預測這個專題後，我慢慢發現資料探勘和統計在思考方式上其實有差別。透過專題，我感受到資料探勘比較像是從資料出發，先觀察可能的規律，再去思考這些結果代表什麼意義，而不是一開始就先有假設去驗證。雖然我在程式實作上參與不多，但透過整理資料、看結果和討論，我對整個分析流程有了比較清楚的感覺，也理解資料探勘和統計的不同之處。整體來說，這學期我對很多技術細節還是不太熟，也不敢說自己真的學會了資料探勘，但至少透過專題和課堂的學習，對資料分析有了初步的理解，也知道這是一個需要反覆思考與觀察的領域。如果未來有機會再接觸相關內容，應該還需要花更多時間去補強。

112029009 林佳

在做實作探勘前，我以為是只要有程式，再將資料丟進去就可以進行分析，但是實際操作起來，除了要查不少資料選擇模型外，資料理解、前處理到結果解讀的完整思考過程，這與我最初的想像有相當大的落差。中間不管是模型選擇還是閾值調整調整程式就調整了不少次。不僅如此，資料處理的先後順序也調整了好幾遍，還需要時刻注意調整是否合理，會不會出現偏頗。一開始讓我覺得挫折的地方是最開始的前處理標準化，因為沒查詢足夠的資料，同時沒注意前面的資料型態，導致使用的標準化錯誤，後面的程序一連串都是不對的。經過期中更改後，在訓練調整模型的部分也不是非常順利，常常出來數值不滿意，或是過擬和，又或者是資料少跑了矩陣或著結果

計數。

不過在經過嘗試和修正，以及負責跑程式組員的提醒後，逐漸開始理解流程，能慢慢去調整，導出良好的結果。我也逐漸理解資料探勘並非追求唯一的標準解，而是根據選擇的研究目的不同從而反覆驗證與解釋的過程。希望這次的經驗能讓我未來能與資料處理共處，做更多的分析嘗試。

112029011 張庭瑜

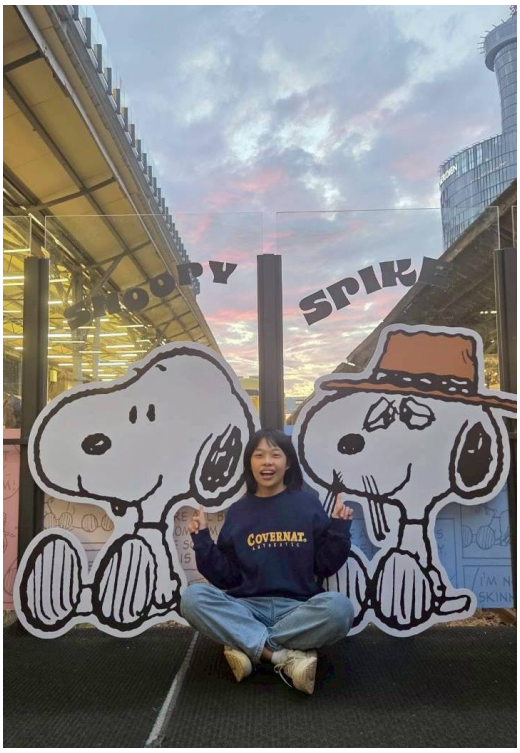
在修完整學期資料探勘的課程後，從原本對機器學習以及模型訓練幾乎都沒有概念，到經過一次又一次的課堂累積基礎概念後，到了學期後段，終於能把這些學到的東西實際應用在自己選的題目上，完成一份完整的資料分析時，雖然過程很燒腦，但也讓人很有成就感。

在分組討論報告的過程中，最常遇到的問題是搞混模型訓練的順序，因為是分工合作，每個人負責的部分都不一樣，有時候會有交接不當的問題，導致有些部分重複做了，有些部分卻漏掉，在協調的過程中就浪費了不少時間。這也讓我意識到，團隊合作中如果能更常追蹤彼此的進度，整個專案進行起來會更有效率。

對於個人來說的話，除了增強了報告的表達能力以外，還讓我擺脫了處理事情過於草率的態度，因為這是一份從頭到尾環環相扣，因此需要紮實去做的專案，只要有一個步驟偷懶或是不全面了解，除了不懂後續的步驟在做什麼，報告被問的時候也會答不上來，那就會很尷尬。

雖然準備報告的時候很辛苦，也和組員一起熬過不少個夜，講了好幾通一講就是兩個小時起跳的電話，但在與他們腦力激盪的過程中，可以更清楚自己的盲點，以及讓邏輯更通順，因此我也覺得收穫了很多。

以下附上我的生活照，希望老師可以認出我。



112029056 林欣璇

回顧這學期的資料探勘課程，如果要用一個詞來形容，那應該是「痛苦與快樂並存」。坦白說，剛開始接觸這門課的時候，我原本以為只是跑程式碼，圖表就會自己跑出來，但現實卻狠狠地打了我一巴掌。

首先，從技術層面來說，我覺得我這學期最大的轉變，在於從盲目地套用模型進階到理解模型在做什麼。我印象最深刻的痛苦時刻，就是在處理糖尿病預測模型的過程。那時候為了處理資料不平衡，我一直在糾結 Borderline-SMOTE 到底要放在哪裡，後來才深刻理解到為了避免資料洩漏，必須在每一折裡面做動態生成，明明是一些微小的細節，卻讓我再跑程式碼的過程中卡關許久。

除了 coding，最讓我崩潰的還有如何定義好壞。以前總覺得準確率越高越好，但真正深入醫學資料後才發現事情沒這麼簡單。在面對不平衡的糖尿病數據時，我必須強迫自己去查閱相關文獻，去理解在醫學統計上，AUC 達到 0.8 以上才算具備良好的鑑別力，以及為什麼在疾病篩檢中 F1-score 往往比單純的 Accuracy 更具參考價值。此外，讓我覺得更具挑戰性的是閾值的決策過程，我們不能無腦地使用預設的 0.5，而是需要找到文獻支持的平衡點。這段過程讓我學會了繪製 Precision-Recall Curve，並試圖在「不漏抓病患」與「不誤判健康人」之間找到最佳解。最終我們依據

數據計算出 0.612 這個閾值，而不是盲目猜測，這不僅讓模型的邏輯更站得住腳，也讓我明白資料探勘的價值不在於跑出一個數字，而在於能不能為這個數字找到強而有力的理論支撐。

最後，專案的最後一哩路——書面報告，也完全不輕鬆。因為這是我第一次當組長，需要在最後將此次書面彙整好，因此需要跟 Word 的分節符號和頁碼搏鬥。為了讓報告呈現最完美的格式，連一個隱形的標點符號都要抓出來。雖然當下覺得「有必要這麼龜毛嗎？」，但看到這份排版整齊、邏輯嚴謹的報告時，我知道這一切都是值得的。

總結這學期的課程，這堂課帶給我的，比我預期的還要多。雖然過程中充滿了挑戰，但現在回頭看，那些與程式碼搏鬥的夜晚、以及為了邏輯正確性而反覆推敲的時刻，其實都轉化成了讓我快速成長的養分。看著最終我們的模型訓練成功，以及手上這份經由我親手整合、排版完美的結案報告，心中的成就感早已超過了過程的疲憊。

附錄:詳細的組員討論與訪談記錄表

欄位	說明
會議編號	例：Meeting #01
日期與時間	2025/12/19 15:00-16:00
地點/形式	實體（CS407）
出席人員	112029002 江信亨 112029008 高翎毓 112029009 林 佳 112029011 張庭瑜 112029056 林欣璇
討論主題	檢討期中報告時發現的問題： 1. 部分欄位（如 Insulin 胰島素、SkinThickness 皮膚厚度）有大量「0」值，探討是否需移除。 2. 插值方法修改，要不要用 KNN
重要決議	1. 針對 Insulin 胰島素進行欄位刪除 2. SkinThickness 皮膚厚度以 KNN 方式補值，其他欄位照舊
待辦事項	112029011 張庭瑜：KNN 補值處理、刪離群值、找相關資料證明做法 112029056 林欣璇：中位數補值、找相關資料證明做法

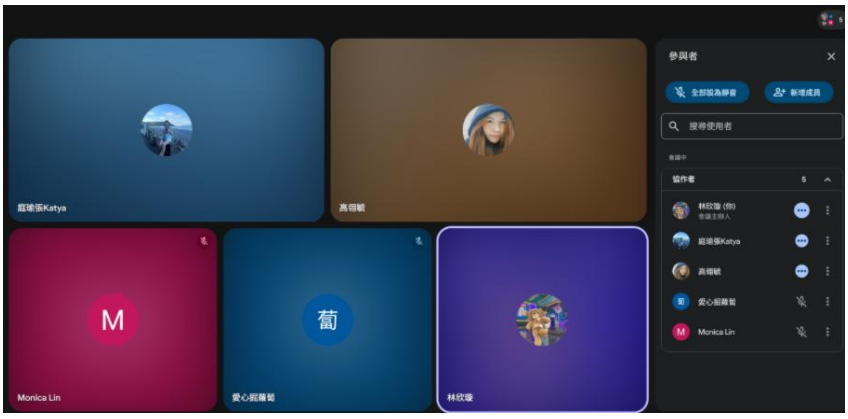
欄位	說明
會議照片	

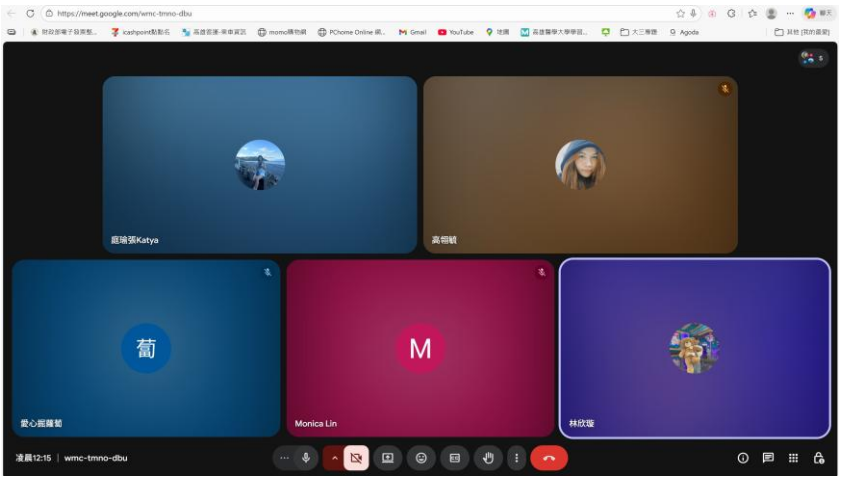
欄位	說明
會議編號	例：Meeting #02
日期與時間	2025/12/26 16:00-18:00
地點/形式	實體（專題教室）
出席人員	112029002 江信亨 112029009 林 佳 112029011 張庭瑜 112029056 林欣璇
討論主題	1. 比較不同演算法在糖尿病資料集上的表現（Random Forest vs. XGBoost）。 2. 討論如何調整參數以避免過度擬合。 3. 認為在醫療預測中，「漏抓病患（偽陰性）」比「誤判健康者（偽陽性）」更嚴重，想調整閾值。
重要決	1. 決定使用 GridSearchCV 來尋找最佳超參數。

欄位	說明
議	2.　了解到閾值調整需要以訓練完模型後，測試集跑完才能調整。
待辦事項	112029009 林 佳：設定超參數並訓練 XGBOOST 模型 112029011 張庭瑜：設定超參數並訓練 XGBOOST 模型 112029056 林欣璇：設定超參數並訓練 Randomforest 模型
會議照片	

欄位	說明
會議編號	例：Meeting #03
日期與時間	2026/01/01 13:00-17:30
地點/形式	線上

欄位	說明
出席人員	112029002 江信亨 112029009 林 佳 112029011 張庭瑜 112029056 林欣璇
討論主題	1. 認為在醫療預測中，「漏抓病患（偽陰性）」比「誤判健康者（偽陽性）」更嚴重，想調整閾值。 2. 是否需要類別平衡
重要決議	1. 最終選用 Borderline SMOTE 進行類別平衡 2. 調整預值測試
待辦事項	112029009 林 佳：XGBOOST 模型以 Borderline SMOTE 進行類別平衡 112029011 張庭瑜：Randomforest 模型以 Borderline SMOTE 進行類別平衡 112029056 林欣璇：調整預值測試、查詢閾值文獻 112029008 高翎毓：PPT 製作
會議照片	

欄位	說明
會議編號	例：Meeting #04 #05
日期與時間	2026/01/07 22:00-00:00 2026/01/08 21:00-00:00
地點/形式	線上
出席人員	112029002 江信亨 112029008 高翎毓 112029009 林 佳 112029011 張庭瑜 112029056 林欣璇
討論主題	1. PPT 製作與練習報告
待辦事項	112029008 高翎毓：PPT 製作
會議照片	

欄位	說明
	 https://meet.google.com/wmc-tmno-dbu YouTube Gmail Agoda Katya M Monica Lin 凌晨12:15 | wmc-tmno-dbu

附錄:上台報告時錄音筆所記下的問題答覆與說明

1.(112029009 林佳)

Q：補 knn 的超參數、距離計算，並且是參考哪些特徵來進行補值？

A：

(一)參考特徵：目標項目-SkinThickness ，參考特徵-Glucose、BloodPressure、BMI、Age、DiabetesPedigreeFunction 以及 SkinThickness

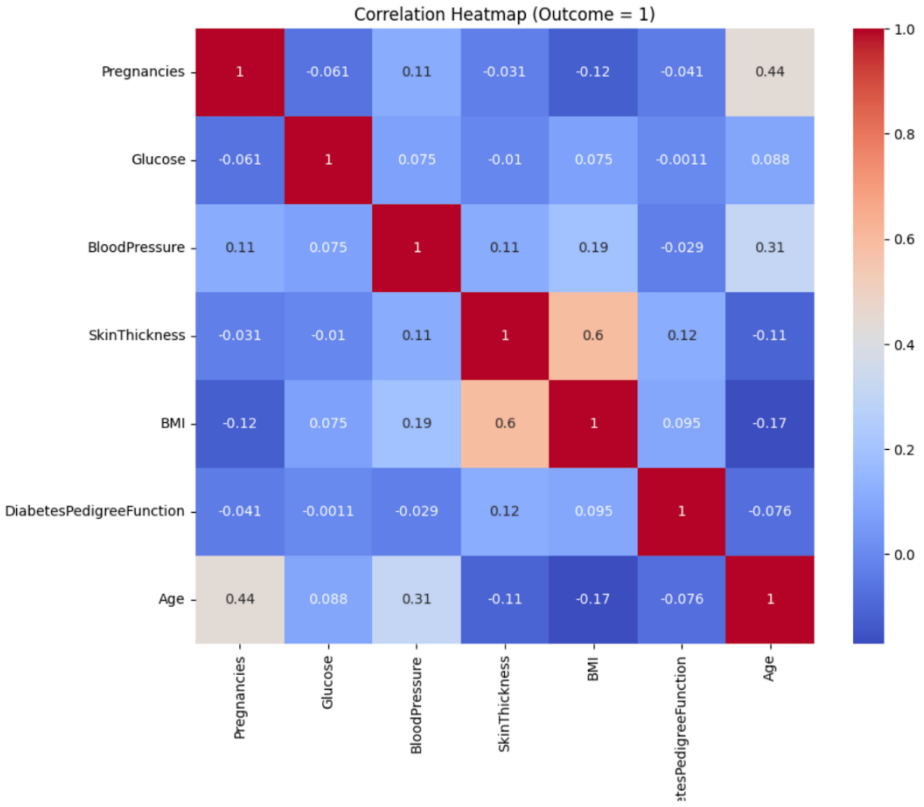
(二)超參數設定：KNN 補值之鄰居數設定為 5，表示每一筆缺失值係由距離最近之 5 筆樣本進行估計。補值時的權重採用預設 uniform weighting，使每一個最近鄰樣本對補值結果具有相同權重，以避免補值結果過度受單一極近鄰樣本影響。

(三)距離計算：以歐式距離進行計算，值前已對所有連續型變數進行 0-1 正規化，確保距離計算過程中各變數具有相同尺度。

補值前，將 SkinThickness 數值為 0 的設為缺失值。因為數值等於 0 並不合理，所以設值為缺失值，避免將其當為真實數值，扭曲距離的抓取。

2.(112029002 江信亨) (112029056 林欣璇)

Q：p.19 熱力圖中 outcome 是類別變數，只能從 outcome=1 的情況下下去看各變數之間的相關係數(圖表分析已寫在資料視覺化那章)



3.(112029008 高翎毓)

(1)交叉結果表格補充 Val 的數據資料

隨機森林Train	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SMOTE前	0=312, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164
SMOTE後	0=312, 1=312	0=313, 1=313	0=313, 1=313	0=313, 1=313	0=313, 1=313
指標 Accuracy	0.9097	0.9308	0.8910	0.9245	0.9015
AUC	0.9874	0.9878	0.9804	0.9837	0.9779
F1	0.8802	0.9054	0.8571	0.8977	0.8691

隨機森林Val	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
set	0=79, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41	
指標 Accuracy	0.7917		0.7395		0.7479		0.7479		0.7563	
AUC	0.8524		0.8018		0.8215		0.8008		0.8399	
F1	0.7253		0.6931		0.6591		0.6809		0.6882	
混淆矩陣 TP FP	33	17	35	25	29	18	32	21	32	20
FN TN	8	62	6	53	12	60	9	57	9	58

XG Boost Train	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SMOTE前	0=312, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164	0=313, 1=164
SMOTE後	0=312, 1=312	0=313, 1=313	0=313, 1=313	0=313, 1=313	0=313, 1=313
指標 Accuracy	0.8638	0.8770	0.8594	0.8658	0.8578
AUC	0.9340	0.9454	0.9287	0.9355	0.9290
F1	0.8686	0.8821	0.8654	0.8679	0.8620

XG Boost Val	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
set	0=79, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41		0=78, 1=41	
指標 Accuracy	0.8333		0.7311		0.7899		0.7479		0.7731	
AUC	0.8740		0.8161		0.8405		0.8124		0.8674	
F1	0.7619		0.6667		0.7126		0.6739		0.6966	
混淆矩陣 TP FP	68	11	55	23	63	15	58	20	61	17
FN TN	9	32	9	32	10	31	10	31	10	31

(2) Q：指標評估有沒有過擬合？沒有的話為什麼？

A：觀察數據變化，可以看到訓練集與測試集的表現非常接近，未出現訓練高、測試低的懸殊落差。這種『成比例』的穩定表現，加上在超參數設定選擇較小的樹深度，證明模型成功找到了血糖與 BMI 等關鍵特徵，因此不具備過擬合。

4. (112029056 林欣璇)

Q：隨機森林每棵樹的數量和放入的特徵有哪些？如果都一樣何來變異？

A：根據我們模型經過 GridSearchCV 最佳化後的結果，我們的隨機森林具體設定如下：

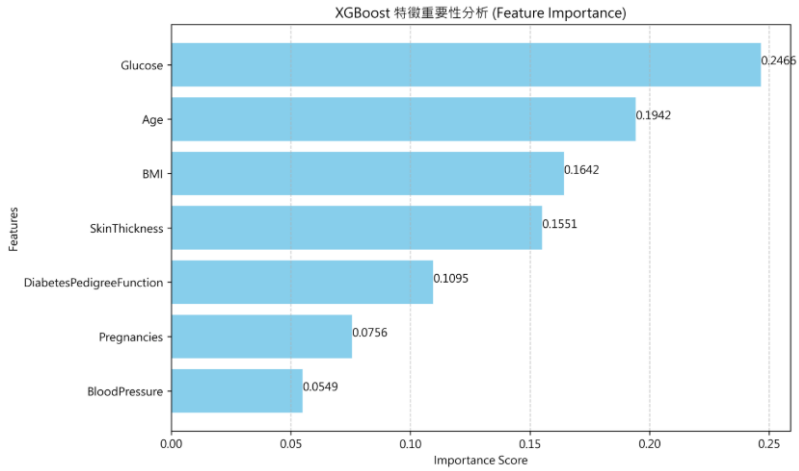
1. 樹的數量 (n_estimators)：模型由 200 棵決策樹組成。
2. 放入的特徵 (max_features)：設定為 'sqrt (開根號) 。

原始資料共有 8 個特徵，則 $\sqrt{8} \approx 2.82$，這意味著，這 200 棵樹在生長過程中的每一個節點 (Node) 要進行分裂時，並不會考慮所有 8 個特徵，而是隨機抽取 3 個特徵作為候選名單，從中挑選最佳的一個來切分。

5. (112029011 張庭瑜)

Q：請問模型可解釋性分析是使用何種 SHAP 來做分析的？

A：原本我們是使用 XGBoost 中內建的模型重要性分析，但後來發現此方法僅能呈現特徵重要與否，卻無法說明該特徵與目標值（Outcome）之間是正相關還是負相關。



由於我們使用的模型為 XGBoost，因此我們決定使用 Tree SHAP 演算法，以確保特徵貢獻度的分配具備數學上的公平性與一致性。。Tree-SHAP 可以跑出兩張圖表，其中一張是 Summary Plot，這張表可以看出所有特徵的影響力排行、與結果的正相關或負相關關係，以及數值大小與 SHAP value 的大小關係。