應用機器學習於糖尿病風險預測：BRFSS資料實證分析

110072250陳冠倫, 109011261謝裕紀, 109034057馮榮晟

前言

糖尿病是一種常見的慢性代謝性疾病，初期症狀不明顯，卻可能帶來嚴重併發症。若能及早發現並介入治療，不僅有助於個人改善生活品質，亦能有效降低整體醫療支出與資源負擔。現行診斷方式多仰賴空腹血糖與糖化血色素（HbA1c）檢測，惟此類抽血檢查在費用、便利性與可及性方面，對資源有限地區仍存實務挑戰。

因此，若能透過簡單、非侵入性的健康指標或問卷資料預測糖尿病風險，將具有高度公共衛生與政策應用價值。本研究即嘗試以機器學習方法建立風險預測模型，支援早期篩檢與健康決策。

本研究使用資料來自美國疾病控制與預防中心（CDC）於 2015 年實施之「行為風險因素監測系統」（Behavioral Risk Factor Surveillance System, BRFSS），該調查透過電話訪問方式蒐集成人健康相關資訊，共計 253,680 筆樣本與 22 個變數。資料涵蓋個人健康狀況、生活習慣與社會經濟背景等，目標變數為 Diabetes_binary，表示是否曾被醫師診斷為糖尿病，為二元分類。共 14 項欄位為二元變數（如 HighBP、Smoker、PhysActivity 等），其餘變數為連續型（如 BMI、Age 等）。

本研究旨在探索機器學習於公共衛生領域之可應用性，建構一套可供民眾使用之糖尿病風險自我預測模型，提升在糖尿病早期發現的可能性。

I. 方法介紹

A. 前處理

此資料集沒有缺失值，而由於可能有同年齡、同狀況但不同人的重複值（Duplicate），因此也不需去除。進行EDA時發現此資料急兩兩變數之間並無超過0.6的相關係數，並且由於多為類別數據，複雜度不高，因此納入全部欄位成為特徵。

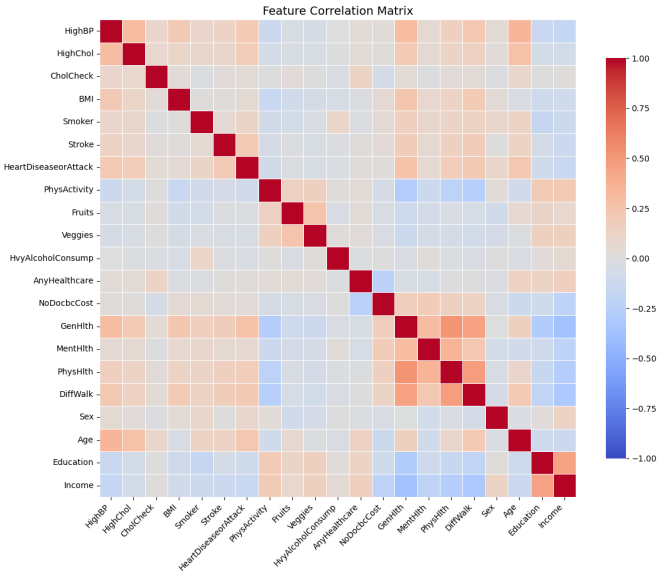

圖1：資料集熱力圖

為提升模型效能與穩定性，本研究於建模前進行一系列資料前處理流程。首先，針對資料集中之 Diabetes_binary 目標變數進行分層抽樣（Stratified Sampling），將資料依正負類別比例切分為訓練集（60%）、驗證集（20%）與測試集（20%）。

接著進行特徵標準化處理。考量原始資料中包含多個二元類別變數與連續數值變數，本研究僅對數值型變數進行標準化處理，類別變數保持原樣。標準化方法採用 Z-score 轉換，將 training set 的數值特徵減去其平均值並除以標準差製成 scaler，使其轉換為平均為 0、標準差為 1 之常態分布。

B 模型介紹

- **Least Square Method**

  本研究首先使用最小平方法（Ordinary Least Squares, OLS）建構線性模型，以預測個體罹患糖尿病的機率後再進行分類任務。先透過訓練資料計算封閉解以取得模型參數θ，公式如下：

$$\theta = (X^\top X)^{-1} X^\top y$$

  其中 X 為標準化後的特徵矩陣，y 為對應之目標變數。預測結果為一連續分數，代表個體傾向於為正類（有糖尿病）之程度。

為強調模型對潛在糖尿病個體的識別能力，本研究將評估指標著重於召回率（Recall）。在公共衛生應用中，錯過高風險個體（False Negative）將可能延誤治療時機，帶來更高的醫療與社會成本，因此提高召回率遠比避免誤報（False Positive）更具實質意義。

為此，我們採用 Fβ 分數（Fβ Score） 作為分類門檻調整的依據，並設 $\beta = 2$，以增加召回率在指標中的權重，同時保留對準確率（Precision）的考量。其數學定義如下所示：

$$F_\beta = (1+\beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

在驗證資料上，我們遍歷不同的 threshold，選擇能最大化 F2 分數之閾值，並將該閾值應用於測試資料的預測中，以進行最終效能檢驗與混淆矩陣（confusion matrix）分析。

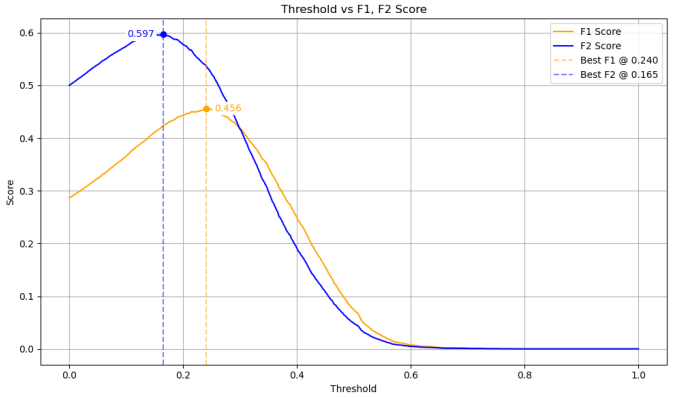

圖2：F1, F2分數在Val set根據threshold的變化 (least square)

- **Naïve Bayes**

作為本研究中第二個建構的預測模型，我們製作 Naïve Bayes 分類器，以機率推論架構建模糖尿病風險，並對應二元分類任務中不同變數類型設計對應之 likelihood function。Naïve Bayes 模型基於條件獨立假設，推導每個樣本 x 在不同類別下的 joint probability，並透過 maximum posterior 作為分類依據：

$$\hat{y} = \arg\max_y P(y|x) = \arg\max_y \frac{P(x|y)\cdot P(y)}{P(x)} \propto P(x|y)\cdot P(y)$$

其中，P(y) 為各類別的先驗機率（prior probability），P(x | y) 則為在特定類別下各特徵的條件機率。因 P(x) 對所有類別固定，可省略不計。

為能同時處理數值型與二元型變數，本研究依據變數類型分別建構如下條件機率：

1. 對於數值型特徵，假設其在給定類別 y 下服從常態分布（normal distribution），其 likelihood 表示如下：

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_{iy}^2}} \exp\left(-\frac{(x_i - \mu_{iy})^2}{2\sigma_{iy}^2}\right)$$

2. 對於二元特徵則採用伯努利分布（Bernoulli distribution）建模，表示如下：

$$P(x_j \mid y) = p_{jy}^{x_j}(1-p_{jy})^{1-x_j}$$

3. 接著將所有 log-likelihood 加總後與 log prior 結合，作為模型之 discriminant function：

$$\log P(x \mid y) = \sum_{i \in \text{numeric}} \log P(x_i \mid y) + \sum_{j \in \text{binary}} \log P(x_j \mid y)$$

$$\hat{y} = \arg\max_y [\log P(y) + \log P(x \mid y)]$$

最後延續先前對線性模型所採用之作法，針對後驗機率調整threshold，並以驗證集找出最大化F2分數的閾值。模型預測時輸出屬於患者的機率值，並據此進行最佳切點分類。

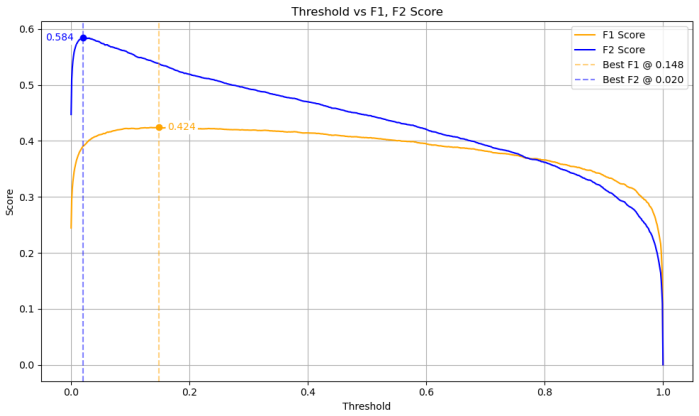

圖3：F1, F2分數在Val set根據threshold的變化 (naïve bayes)

- **Logistic Gradient Descent**

本研究亦實作一套基於梯度下降法（Gradient Descent）之邏輯斯回歸（Logistic Regression）模型，作為具備解釋性且具泛化能力的機率預測方法。模型以 binary cross-entropy 作為損失函數（loss function），也就是負對數似然函數（negative log-likelihood function），公式如下：

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)], \quad \hat{p}_i = \sigma(\mathbf{x}_i^\top \mathbf{w} + b)$$

其中，σ(·) 為 sigmoid 函數，用於將線性組合映射為機率：

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

參數更新以 batch 梯度下降法進行，每次迭代更新偏差項 b 與權重向量 w，學習率經實驗後設定為 0.01，最多訓練 10,000 次。為避免過擬合並提升收斂效率，我們於訓練期間實作 early stopping 機制，當驗證集連續 100 epoch 的 val loss 變化小於 0.0001，即代表損失下降已達平坦，提前終止訓練並保留表現最佳之參數組。訓練過程中亦記錄損失值，並繪製訓練與驗證損失曲線（loss curve）以觀察收斂趨勢與過擬合情況。

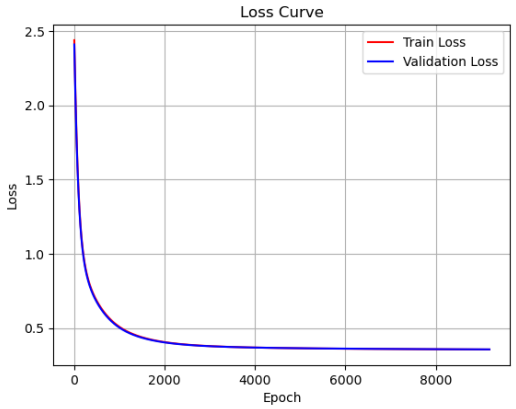

圖 4：Logistic Regression 模型訓練與驗證損失曲線

在模型評估階段，預測結果為每筆樣本屬於患者的機率值，並進一步以驗證集為基礎搜尋最佳 threshold。我們同樣 F2 分數作為主要指標，調整機率閾值以最大化在 recall 權重較高場景下的分類效能。

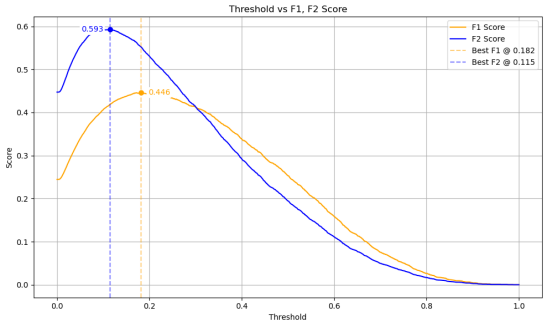

圖5：F1, F2分數在Val set根據threshold的變化 (logistic GD)

- **Undersampling 後重新建模**

為探討類別不平衡對模型造成之影響，我們設計另一組訓練流程，將訓練集與驗證集分別進行 1:1 欠抽樣（undersampling），即隨機保留與正類樣本等量的負類樣本，形成平衡樣本集。資料標準化程序仍以 undersample 後進行 Z-score 轉換。

Undersampling 後，我們重新訓練 Logistic Regression 模型，並重複前述 threshold 搜尋與效能評估程序，藉此觀察平衡資料對 recall 與 precision 表現的影響。最終，我們比較欠抽樣與未抽樣模型在 test set 上的 F2 分數與混淆矩陣結構，以提供實務上不同取捨下的參考依據。

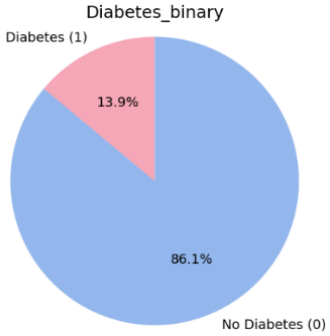

圖 6：目標類別不平衡

- 其他方法

為求提升模型效能與解釋能力，本研究亦曾嘗試其他資料前處理與降維技術，如主成分分析（PCA）與線性判別分析（LDA），以簡化特徵空間並降低模型複雜度。然而在本資料集中，原始變數已具良好辨識度，經降維後反而導致部分資訊流失，影響分類效能，故未納入主要分析流程。

此外，亦實作基於 Gradient Descent 的 Linear Regression 作為二元分類應用，惟預測效能略低於 Logistic Regression，故最終僅保留後者作為主要線性模型呈現。

整體而言，本研究在模型選擇上，以效能、解釋性與訓練穩定性為主要考量，並於多項嘗試後保留三組表現最為穩定且具公共衛生應用潛力之方法進行深入討論。

II. 結果與討論

	Least Square	Naïve Bayes	Logistic GD	Undersample Logistic GD
TN	27,844	25,230	28,924	16,686
FP	15,824	18,438	14,744	26,982
TP	5,965	6,119	5,740	6,814
FN	1,105	951	1330	256
Precision	0.274	0.249	0.280	0.202
Recall	0.844	0.865	0.812	0.964
F2 score	0.596	0.579	0.588	0.549

表 1：四種模型的各項指標與數字

從表 1 可觀察各模型在測試資料上的預測表現，可進一步比較其在 Precision、Recall 與 F2 Score 三項分類指標的權衡與差異。

首先，Least Square 模型雖然原理上為回歸方法，但透過對預測值調整閾值（threshold）後，其在分類任務中表現仍具參考價值。該模型於測試集上的 Recall 達 0.844，能有效識別大部分糖尿病個體，同時 Precision 為 0.274，F2 分數為 0.596，為所有模型中最高。儘管其預測架構簡單，卻在 Recall 與 Precision 間取得良好平衡，適合用於早期篩檢階段。

接著觀察 Naïve Bayes 模型，其假設特徵彼此獨立，並同時處理連續型與二元變數。該模型達成了 最高的 Recall（0.865），展現其在識別正類（糖尿病）樣本上的積極性。然而，Precision 僅為 0.249，表示其亦產生了大量的誤判（FP 數達 18,438），反映模型傾向過度預測為正類，可能導致不必要的醫療檢測與成本。其整體 F2 分數為 0.579，低於 Least Square 與 Logistic GD 模型，顯示其在 Recall 提升下有所折衷。

Logistic 模型經 Gradient Descent 訓練後，在測試資料中取得 Precision 為 0.280、Recall 為 0.812 的結果，是四種模型中 Precision 最高者。雖然 Recall 僅居第三，但在 Precision 與 Recall 之間取得相對穩定的平衡，使得其 F2 分數為 0.588，整體分類表現亦不俗。該模型適合在需要兼顧誤診與漏診風險的應用場景，展現不錯的泛化能力。

最後，在進行 undersampling 後的 Logistic Regression 模型（將訓練與驗證資料中 0/1 樣本比例調整為 1:1），其 Recall 大幅提升至 0.964，成功識別出幾乎所有糖尿病個體（FN 僅剩 256 筆），展現極強的正類召回能力。但也因此造成正類預測的極端偏重，產生了高達 26,982 筆的假陽性（FP），Precision 下降至 0.202，為四種模型中最低。整體 F2 分數為 0.549，低於其他模型，反映出在極端提升 Recall 的同時，需付出相當代價。

綜合而言，若應用情境為初步篩檢，Recall 與 F2 分數是優先考量指標，則應選擇 Least Squar 或 Logistic Regression 模型，能兼顧識別力與誤判風險。若目標為最大限度避免漏診，則 undersampled Logistic Regression 雖有誤判代價，仍具臨床意義。但如強調精準預測、避免誤診成本，則需針對 threshold 做進一步微調或結合其他策略（如校正機率、後續篩檢階段設計）以達成更佳平衡。

III. 結論

本研究成功以多種機器學習方法建構糖尿病風險預測模型，驗證簡單問卷與基本健康指標即可有效識別高風險個體的可行性。在未使用抽血檢查等高成本手段下，模型已展現良好效能，尤其在 Recall 與 F2 分數方面，顯示其具備早期篩檢與公共衛生應用潛力。

從模型表現來看，Least Square 與 Logistic Regression（GD）兩者在召回率與 F2 分數間取得良好平衡，適合作為大規模初篩工具；Naïve Bayes 雖 Recall 最高，但誤判風險偏高；而透過 undersampling 優化 Recall 的 Logistic Regression 雖幾近全數識別高風險個體，卻伴隨顯著精確率損失，適合應用於「寧濫勿漏」的高敏感性場景。不同模型策略間的取捨，亦反映出公共衛生決策中對「誤判」與「漏診」風險的平衡關鍵。

本研究已展示以問卷與健康指標建構糖尿病預測模型的可行性，並在召回率導向下取得良好分類效能。未來可進一步引入非參數式機器學習方法，以突破傳統模型在資料分布假設上的限制，進一步提升模型在複雜特徵結構下的彈性與準確性，擴大實務應用的潛力。

整體而言，本研究展示了「低成本、高可及性」預測模型於疾病防治上的巨大潛力，未來可應用於行動端自評工具、社區初步篩檢與政策風險評估，進一步補足現行醫療體系之不足，落實預防醫學之核心精神。

參考資料

- 如何做undersampling
  https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/

- 在資料不平衡的狀況下如何調整閾值
  https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/?utm_source=chatgpt.com

- F beta Score定義與使用方法
  https://blog.csdn.net/Libo_Learner/article/details/83615715