

淡江大學統計學系數據科學碩士班

R程式設計 期末論文

比較機器學習方法在中風預測之應用

整合年齡分層與風險因子分析的預測

模型研究

— 以GAM、隨機森林與XGBoost進行實證分析

指導教授：唐嘉宏 教授

學生：施宇鴻、鄭傑丞

學號：613890218、613890069

中華民國 113 年 12 月

目錄

第一章 緒論	7
第一節 研究背景與動機	7
第二節 研究目的	8
一、 建立多元預測模型	8
二、 評估關鍵風險因子	8
三、 模型效能評估與比較	9
四、 實務應用價值	9
第三節 研究範疇與限制	9
一、 研究範疇	9
二、 研究限制	10
三、 研究效度考量	11
第二章 研究方法	11
第一節 數據來源與處理	11
第二節 預測模型的選擇與設計	12
一、 資料分割	12
二、 模型設計	13
三、 模型評估指標	14
第三節 效能評估指標	15
一、 ROC曲線與AUC值	15
二、 敏感度與特異度分析	16
三、 精確度與召回率	17
四、 模型訓練與測試框架	17
五、 模型比較準則	18
六、 效能評估流程	18
第三章 研究結果與分析	19
第一節 敘述統計	20
一、 基本人口統計特徵	20
二、 健康狀況指標	21

三、生活型態特徵.....	24
四、中風發生率	25
第二節 年齡分層風險分析	26
一、年齡層別與中風發生率分析	27
二、年齡與其他風險因子的交互作用	28
三、性別與其他風險因子的交互作用	30
四、中風對各類別因子關聯性分析	33
五、關鍵風險因子組合分析	35
第三節 預測模型建立與評估	36
一、模型參數設定與訓練過程.....	36
二、模型預測效能比較.....	40
第四節 模型預測準確度分析	43
一、預測結果的穩定性分析	44
二、穩定性指標評估.....	45
第五節 模型在不同族群中的預測表現	47
一、年齡層別預測效能.....	48
第四章 結論與建議.....	50
第一節 研究結論.....	50
一、中風預測的成果與貢獻	50
第二節 研究建議.....	51
一、未來研究建議.....	51
二、臨床實務中的可能應用	52
參考文獻.....	53

表目錄

表 1性別與風險因子之間的卡方檢定表	31
表 2不同性別的健康風險因子比率表	32
表 3不同性別與吸菸狀態的中風比率表	32
表 4中風與各類別之間的卡方檢定表	34
表 5中風與各類別之間的T檢定表	35
表 6三種模型預測效能評估	41

圖目錄

圖 1性別分佈圓餅圖	20
圖 2年齡分佈直方圖	21
圖 3高血壓狀況與年齡分佈長條圖	21
圖 4心臟病患者分佈圓餅圖	22
圖 5平均血糖濃度分佈長條圖	23
圖 6 BMI 身體質量指數分佈長條圖	23
圖 7吸菸狀況分佈圓餅圖	24
圖 8居住型態分佈圓餅圖	25
圖 9中風發生率分佈圓餅圖	26
圖 10年齡層別與中風發生率長條圖	28
圖 11不同年齡層中風及風險因子發生率長條圖	29
圖 12不同年齡層中風及風險因子發生率折線圖	29
圖 13不同年齡層的平均血糖水平、BMI及中風發生率	30
圖 14不同性別的健康風險因子比較長條圖	32
圖 15不同性別與吸菸狀態的中風比率長條圖	33
圖 16風險因子組合與中風發生率長條圖	36
圖 17 ROC曲線比較模型的性能分析圖	45
圖 18模型在不同閾值下的敏感度與特異度表現比較	47
圖 19不同年齡層模型預測性能比較分析長條圖	50

方程式目錄

方程式 1廣義加成模型.....	37
方程式 2隨機森林模型.....	38
方程式 3XGBOOST模型.....	39
方程式 4目標函數公式.....	39

第一章 緒論

第一節 研究背景與動機

腦中風(stroke)是全球主要的死亡和失能原因之一，對公共衛生造成重大影響。根據世界衛生組織(WHO)的最新統計資料顯示，全球每年約有1500萬人罹患中風，其中500萬人因此死亡，另有500萬人導致永久失能。在台灣，中風不僅位居十大死因之一，更為家庭和社會帶來沉重的醫療與長期照護負擔。

世界衛生組織將中風定義為24小時以上的腦神經功能缺損，或在24小時內死亡的狀況。隨著現代醫療技術的進步，中風的治療與預防策略不斷革新，然而研究指出，首次中風病患在5年內仍有高達10%的機率發生二次中風。此一數據凸顯了建立有效預測模型與風險評估機制的迫切性。

近年來，隨著大數據分析與機器學習技術的蓬勃發展，運用統計方法建立精確的中風預測模型逐漸成為可能。本研究採用多元分析方法，結合三種不同特性的統計模型：傳統的廣義可加模型(Generalized Additive Model, GAM)、機器學習中的隨機森林(Random Forest)，以及當代備受關注的極限梯度提升(eXtreme Gradient Boosting, XGBoost)演算法，致力於建立更準確的中風預測模型。

本研究具體著重於三個面向：

1. 年齡分層分析：系統性探討不同年齡層的中風風險特徵與發展模式

2. 風險因子交互作用：深入研究高血壓、心臟病等主要危險因子之間的相互影響關係
3. 預測模型比較：透過多面向指標評估不同機器學習方法在中風預測上的表現

透過建立準確的預測模型，可協助醫療人員更有效地識別高風險族群，及早實施預防性措施，降低中風發生的機率。此外，精確的風險評估也有助於醫療資源的最優配置，進而減輕整體社會的醫療負擔。

第二節 研究目的

本研究旨在透過統計分析方法，建立可靠的中風預測模型，並深入探討影響中風發生的關鍵因素。具體研究目的如下：

一、 建立多元預測模型

本研究運用三種不同特性的統計模型：廣義可加模型(GAM)、隨機森林(Random Forest)及極限梯度提升(XGBoost)演算法，分析並比較其在中風預測上的表現。透過這些先進的統計方法，我們期望能建立更精確的預測模型，提供臨床實務更有力的決策支援工具。

二、 評估關鍵風險因子

本研究著重於四個主要預測變數的分析：

1. 平均血糖濃度(Average Glucose Level)：評估血液中葡萄糖含量對中風風險的影響。
2. 高血壓(Hypertension)：探討高血壓病史與中風發生的關聯性。

3. 心臟病(Heart Disease)：分析心臟病史對中風風險的影響程度。
4. 年齡(Age)：研究不同年齡層的中風風險分布。

三、 模型效能評估與比較

本研究將透過以下方法評估模型效能：

1. 建立受試者操作特徵(ROC)曲線，評估各模型的分類準確度
2. 計算曲線下面積(AUC)，量化並比較不同模型的預測能力
3. 分析各模型的敏感度(Sensitivity)和特異度(Specificity)，全面評估預測效能

四、 實務應用價值

本研究最終目標是將研究成果轉化為實務應用工具，協助：

1. 醫療人員進行更準確的風險評估
2. 制定個人化的預防策略
3. 優化醫療資源的配置效率

透過這些研究目的的實現，我們期望能為中風預防和風險管理提供更有效的科學依據，進而降低中風對個人和社會的衝擊。

第三節 研究範疇與限制

一、 研究範疇

1. 數據來源：

本研究使用的數據集包含多個醫療機構收集的病患資料，涵蓋了各年齡層的個案資訊。數據集中包含多項生理指標和健康狀況變數，為確保分析的準確性，我們採用完整的觀察值進行研究。

2. 分析範圍

本研究主要聚焦於以下面向：

- 生理指標：平均血糖濃度、體重指數(BMI)
- 健康狀況：高血壓、心臟病史
- 人口特徵：年齡、性別
- 生活習慣：吸菸狀態

3. 研究時間

數據收集期間涵蓋多個時間點的觀察值，提供了較為全面的時序性分析基礎。

二、 研究限制

1. 數據來源：

- 部分變數可能存在測量誤差
- 某些重要的潛在風險因子可能未被納入數據集
- 數據收集過程中可能存在選擇性偏誤

2. 方法限制

- 模型預測結果仍需經過臨床驗證
- 預測模型可能無法完全捕捉所有影響中風的複雜因素
- 模型的泛化能力可能受限於訓練數據的代表性

3. 應用限制

- 研究結果主要適用於類似特徵的族群
- 預測模型需要定期更新與校正
- 臨床應用時需要配合專業醫療判斷

三、研究效度考量

1. 內部效度

- 確保數據的品質與完整性
- 採用適當的統計方法處理遺漏值
- 進行模型診斷與驗證

2. 外部效度

- 考慮研究結果的推論範圍
- 評估模型在不同族群中的適用性
- 檢視結果的實務應用可行性

第二章 研究方法

第一節 數據來源與處理

本研究使用 Fedesoriano 於 2020 年在 Kaggle 平台發布的腦中風預測資料集（Stroke Prediction Dataset）[1]作為主要分析數據。在數據前處理的第一步驟，我們移除了不具分析意義的識別碼（X.id）欄位。接著，針對體重指數（BMI）欄位中的遺失值進行處理，考量到BMI為關鍵的健康指標，本研究採用完整案例分析法（complete case analysis），將含有BMI遺失值的觀測值予以剔除。

為優化分析品質，我們進一步清理類別變數。在性別（gender）欄位中，由於標示為「其他」（other）的樣本數量過少，可能影響統計分析的穩定性，因此予以移除。同時，在吸菸狀態（smoking_status）欄位中，標示為「未知」（unknown）的觀測值因缺乏明確資訊，亦予以刪除。

經過上述資料清理流程，本研究最終保留了3,425筆完整觀測值及11個變數欄位。資料集中保留的變數包括性別（gender）、年齡（age）、高血壓（hypertension）、心臟病（heart_disease）、婚姻狀況（ever_married）、工作類型（work_type）、居住類型（Residence_type）、平均血糖值（avg_glucose_level）、體重指數（bmi）、吸菸狀態（smoking_status）以及中風（stroke）等重要指標。此資料規模足以支持後續的統計分析與預測模型建構。

第二節 預測模型的選擇與設計

本研究採用三種不同的機器學習模型進行中風預測分析，以探討不同演算法在處理醫療預測問題上的表現差異。選擇的模型包括：能夠捕捉非線性關係的廣義加性模型（Generalized Additive Model, GAM）、具有強大分類能力的隨機森林（Random Forest）以及在預測競賽中表現優異的極限梯度提升（XGBoost）。

一、資料分割

為確保模型評估的客觀性，本研究採用隨機抽樣的方式將資料集分割為訓練集（Training Set）和測試集（Test Set），比例為7:3。為確保實驗結果的可重複性，設定隨機種子（Random Seed）為1035。在資料分割過程中，我們保持各類別的分布比例，以避免樣本分配不均衡影響模型訓練效果。

二、 模型設計

1. 廣義加性模型（GAM）

GAM模型的設計考慮了變數間的非線性關係，模型構建如下：

- 連續變數：使用平滑函數 $s()$ 處理年齡（age）、平均血糖值（avg_glucose_level）和體重指數（bmi），以捕捉這些變數與中風風險間的非線性關係
- 類別變數：直接納入高血壓（hypertension）、心臟病（heart_disease）和性別（gender）
- 採用二項式家族（binomial family）配合邏輯連結函數（logit link function）進行建模

2. 隨機森林（Random Forest）

隨機森林模型的主要參數設定包括：

- 決策樹數量（ntree）：設定為500，以確保模型的穩定性
- 預測變數：包含平均血糖值、體重指數、年齡、高血壓和心臟病
- 採用分類模式（classification）進行預測
- 每個節點的最小樣本數設定為預設值，以平衡模型的複雜度和準確性

3. 極限梯度提升（XGBoost）

XGBoost模型的具體參數設定如下：

- 目標函數：採用二元分類的邏輯回歸（binary:logistic）
評估指標：使用AUC（Area Under Curve）

- 學習率 (eta)：設定為0.1，以達到較好的收斂效果
- 最大樹深度 (max_depth)：設定為6，避免過度擬合
- 最小子節點權重 (min_child_weight)：設定為1
- 訓練迭代次數 (nrounds)：設定為100，並配合早期停止機制
- 早期停止回合數 (early_stopping_rounds)：設定為10，防止過度擬合

三、模型評估指標

為全面評估模型性能，本研究採用多項評估指標：

1. 準確率 (Accuracy)：

- 計算公式： $(\text{真陽性} + \text{真陰性}) / (\text{總樣本數})$
- 評估模型整體預測準確程度

2. 靈敏度 (Sensitivity)：

- 計算公式： $\text{真陽性} / (\text{真陽性} + \text{假陰性})$
- 評估模型正確識別中風病例的能力

3. 特異度 (Specificity)：

- 計算公式： $\text{真陰性} / (\text{真陰性} + \text{假陽性})$
- 評估模型正確識別非中風案例的能力

4. F1分數 (F1 Score)：

- 計算公式： $2 \times (\text{精確率} \times \text{召回率}) / (\text{精確率} + \text{召回率})$
- 綜合評估模型的精確率和召回率

5. ROC曲線下面積（AUC-ROC）：

- 反映模型在不同閾值下的分類性能
- 評估模型的整體鑑別能力

本研究基於醫療預測的需求，選擇廣義加性模型（GAM）、隨機森林（Random Forest）與極限梯度提升（XGBoost）三種具有代表性的機器學習模型，並依據其特性進行設計與優化，以探討不同算法在中風預測上的表現差異。通過合理的資料分割和多維度的評估指標（如準確率、靈敏度、特異度、F1分數及AUC-ROC），我們確保了分析結果的客觀性與全面性，為後續中風風險的準確評估提供了一個基礎準則。

第三節 效能評估指標

本研究採用全面的效能評估體系，透過多元指標綜合評估預測模型的性能，確保研究結果的可靠性與實用價值。主要評估指標包含以下幾個關鍵面向：

一、 ROC曲線與AUC值

1. ROC曲線

接收者操作特徵曲線（Receiver Operating Characteristic Curve, ROC）是評估二元分類模型性能的重要工具。其基本原理為：

- 橫軸：假陽性率（False Positive Rate, FPR）= $FP / (FP + TN)$
- 縱軸：真陽性率（True Positive Rate, TPR）= $TP / (TP + FN)$
- 透過調整分類閾值，繪製出完整的ROC曲線

2. AUC值解釋

曲線下面積（Area Under Curve, AUC）提供了模型表現的量化指標：

- $AUC = 1$ ：完美分類
- $0.9 < AUC < 1$ ：優秀
- $0.8 < AUC \leq 0.9$ ：良好
- $0.7 < AUC \leq 0.8$ ：一般
- $0.6 < AUC \leq 0.7$ ：勉強
- $AUC \leq 0.6$ ：不及格

二、 敏感度與特異度分析

1. 敏感度（Sensitivity）

定義：正確識別出實際陽性案例的比率

計算公式： $Sensitivity = TP / (TP + FN)$

臨床意義：

- 反映模型發現潛在中風患者的能力
- 降低漏診風險
- 提早及早干預的可能性

2. 特異度（Specificity）

定義：正確識別出實際陰性案例的比率

計算公式： $Specificity = TN / (TN + FP)$

臨床意義：

- 避免過度診斷

➤ 減少不必要的醫療資源消耗

➤ 降低患者心理負擔

3. 權衡考量

➤ 敏感度與特異度的平衡

➤ 根據臨床需求調整閾值

➤ 考慮誤診成本與漏診風險

三、精確度與召回率

1. 精確度

定義：預測為陽性中實際為陽性的比率

計算公式： $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$

應用價值：評估預測的可靠性

2. F1 score

定義：精確度與召回率的調和平均

計算公式： $\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$

綜合評估指標：平衡精確度與召回率

四、模型訓練與測試框架

1. 資料分割策略

訓練集（70%）：用於模型訓練與參數調優 測試集（30%）：用於性能評估 採用分層抽樣確保資料分布一致性

2. 交叉驗證方法

- 採用k-fold交叉驗證 (k=5)
- 評估模型穩定性
- 防止過擬合現象

3. 參數優化

- 網格搜尋 (Grid Search)
- 貝葉斯優化 (Bayesian Optimization)
- 基於驗證集表現選擇最優參數

五、 模型比較準則

1. 綜合評分體系

- 基礎指標：準確率、精確度、召回率
- 平衡分數： $(\text{敏感度} + \text{特異度}) / 2$
- 綜合得分： $(\text{AUC} + \text{平衡分數}) / 2$

2. 臨床實用性評估

- 預測速度
- 模型解釋性
- 實施成本
- 維護難度

六、 效能評估流程

1. 初步評估

- 基礎模型訓練

- 參數優化

- 交叉驗證

2. 深入驗證

- 獨立測試集評估

- 性能指標計算

- 穩定性分析

3. 比較分析

- 多模型性能對比

- 統計顯著性檢驗

- 綜合評分計算

透過這套系統化的評估框架，我們能夠全面評估各個預測模型的性能，為模型選擇提供客觀依據。該評估體系不僅確保了研究的科學性，也為臨床應用提供了可靠的理論支持。

第三章 研究結果與分析

本研究使用的中風預測資料集包含3,425筆觀察值，涵蓋多個重要的健康指標和人口統計特徵。以下將詳細分析各變數的分布特性及其統計特徵。

第一節 敘述統計

一、基本人口統計特徵

1. 性別分布

性別分布顯示如下圖（圖一）所示，女性樣本共有2,086人（60.9%），男性樣本共有1,339人（39.1%）。女性比例略高於男性，但整體性別分布仍保持在合理範圍內。

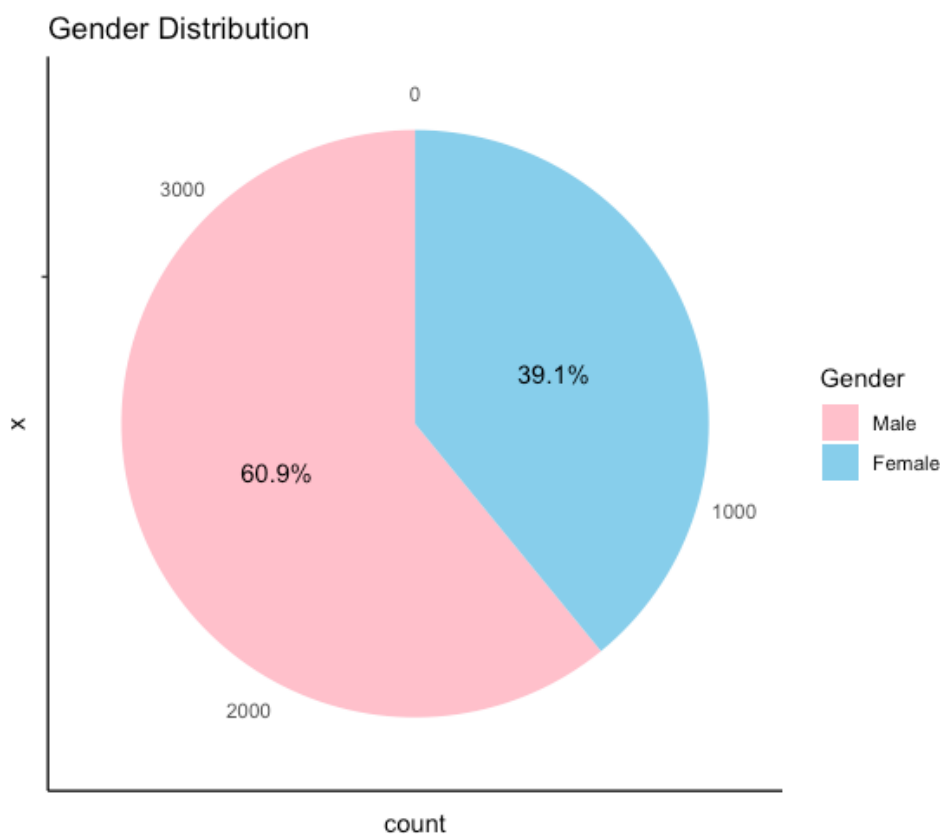


圖 1性別分佈圓餅圖

2. 年齡分布

年齡分布如下圖（圖二）所示涵蓋從10歲到82歲，最小值為10歲，最大值為82歲，平均年齡為48.65歲，中位數為50歲。整體年齡分布呈現良好的跨度，涵蓋青少年到老年的各年齡層，而中位數與平均數相近，顯示樣本的年齡分布相對均衡。

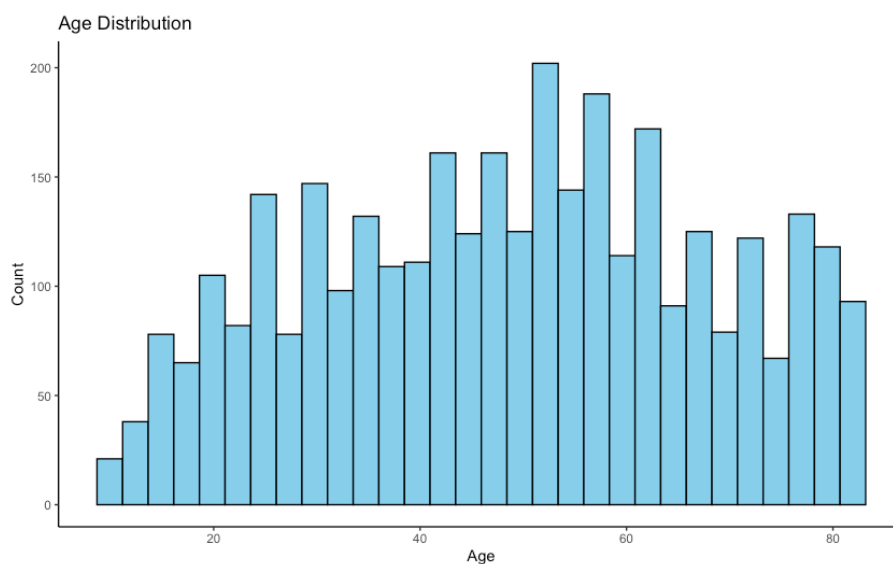


圖 2年齡分佈直方圖

二、 健康狀況指標

1. 高血壓狀況

如下圖（圖三）所示，樣本中無高血壓者共有3,017人（88.1%），有高血壓者共有408人（11.9%）。資料顯示大多數樣本為無高血壓者，少數樣本為有高血壓者。

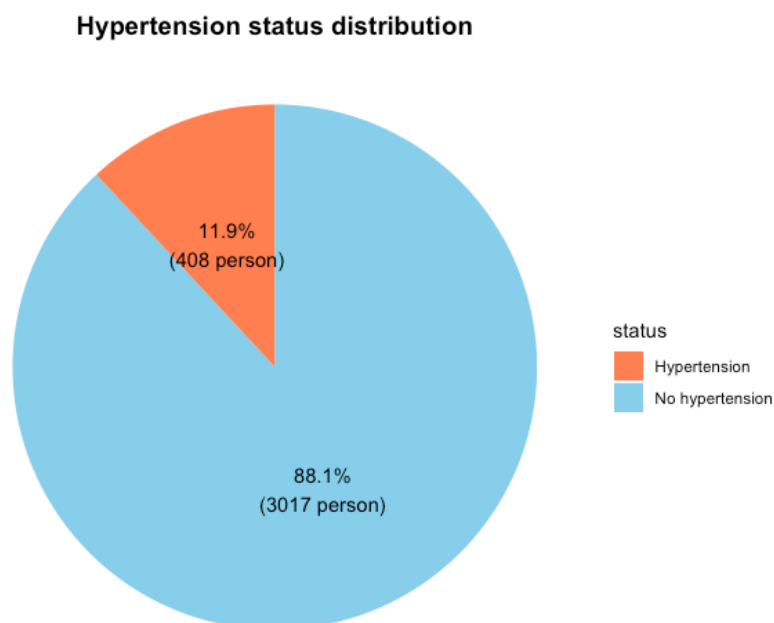


圖 3高血壓狀況與年齡分佈長條圖

2. 心臟病史

如下圖（圖四）所示，樣本中無心臟病者共有3,219人（94%），有心臟病者共有206人（6%）。資料顯示絕大多數樣本無心臟病，僅有少部分樣本為心臟病患者。

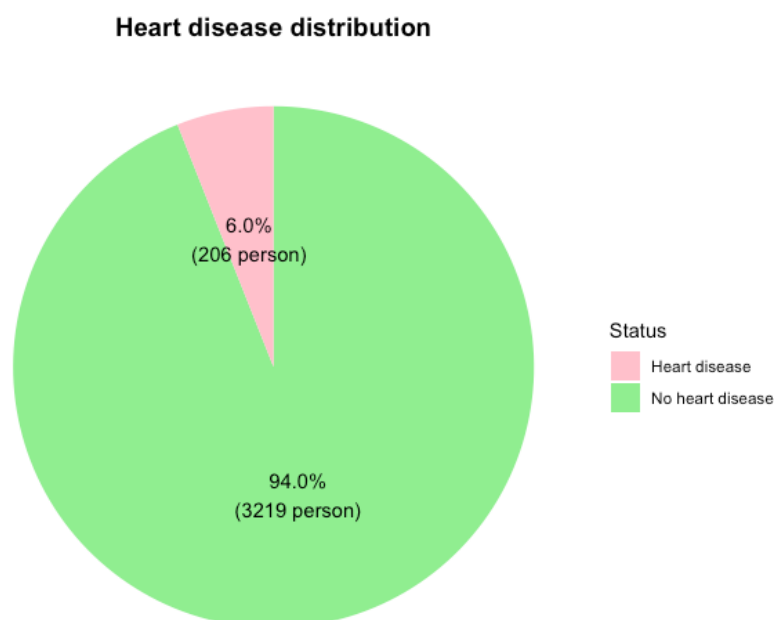


圖 4心臟病患者分佈圓餅圖

3. 平均血糖濃度

平均血糖濃度的分佈情形如下圖（圖五）所示，其中最小值及最大值分別為 55.12 mg/dL 與 271.74 mg/dL，平均血糖濃度值為108.31 mg/dL，中位數為92.35 mg/dL。整體分布呈現右偏，顯示部分樣本的血糖濃度值較高，對平均值產生一定的拉升影響。

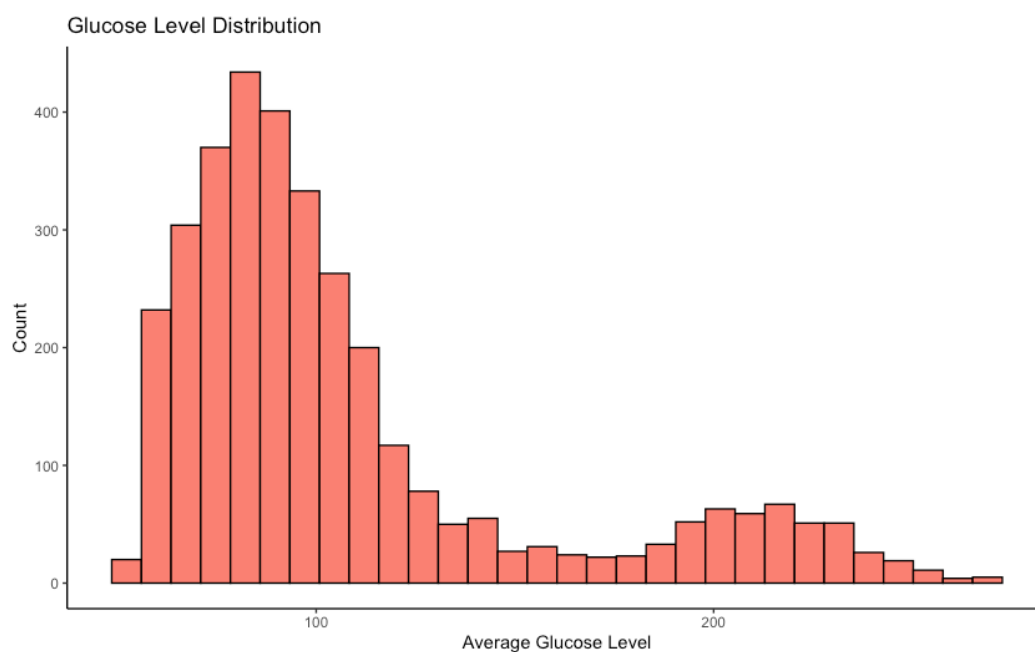


圖 5平均血糖濃度分佈長條圖

4. 身體質量指數 (BMI)

身體質量指數的分佈情形如下圖（圖六）所示，BMI的最小值為11.50，最大值為92.00，平均值為30.29，中位數為29.10。平均值超過30，顯示樣本中存在相當比例的過重或肥胖個案，反映出健康風險的潛在分布特徵。

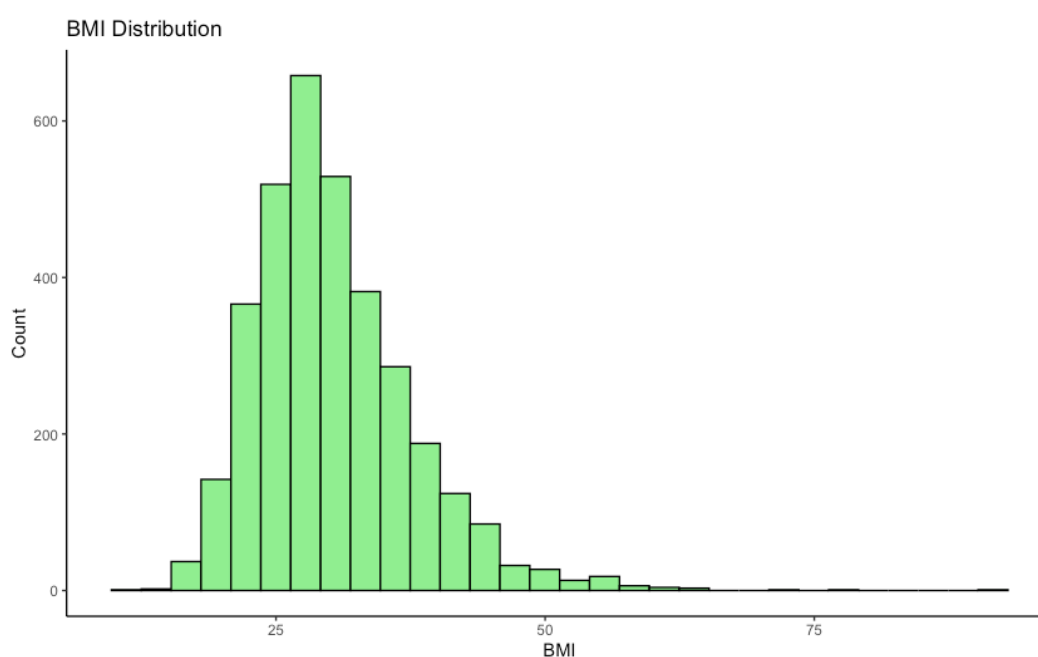


圖 6 BMI 身體質量指數分佈長條圖

三、生活型態特徵

1. 吸菸狀況

吸菸狀況如下圖（圖七）所示，樣本中從未吸菸者共有1,852人（54.1%），曾經吸菸者為836人（24.4%），目前吸菸者為737人（21.5%）。資料顯示多數樣本為從未吸菸者，曾經吸菸與目前吸菸者比例相對接近，但總體吸菸相關者占比仍達近半數。

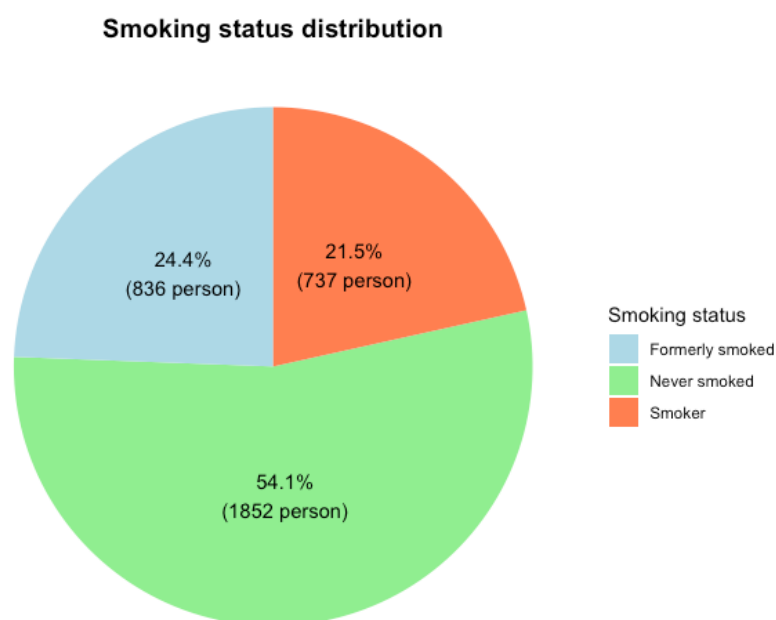


圖 7吸菸狀況分佈圓餅圖

2. 居住類型

樣本中居住類型如（圖八）所示，居住於農村地區的人數為1,680人（49.1%），居住於城市地區的人數為1,745人（50.9%）。資料顯示農村與城市地區的分布相對均衡，兩者人數比例接近各半。

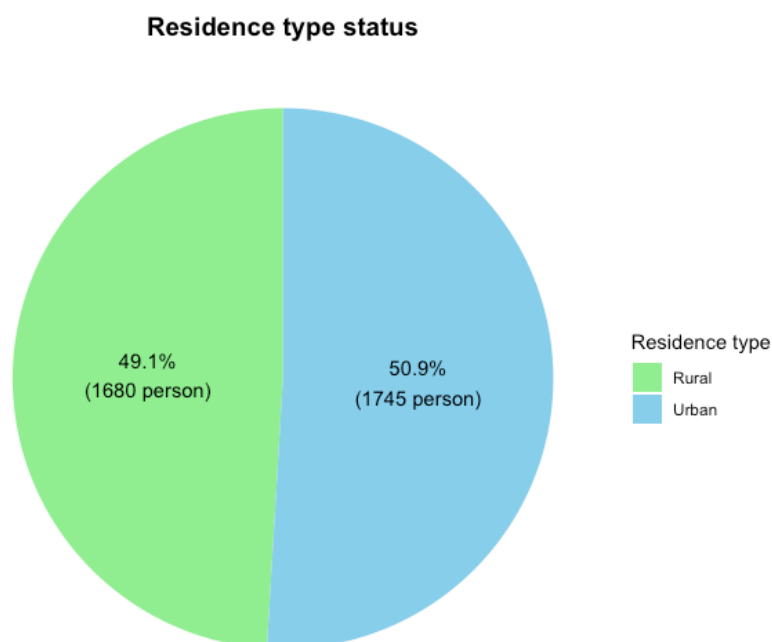


圖 8居住型態分佈圓餅圖

四、 中風發生率

整體樣本的中風發生率分佈如（圖九）所示，未發生中風者共有3,245人（94.7%），發生中風者為180人（5.3%）。這種不平衡的數據分布反映了中風案例在現實中相對罕見的特性，同時也提醒我們在後續建模時需特別關注類別不平衡的處理，以確保模型的預測效果。

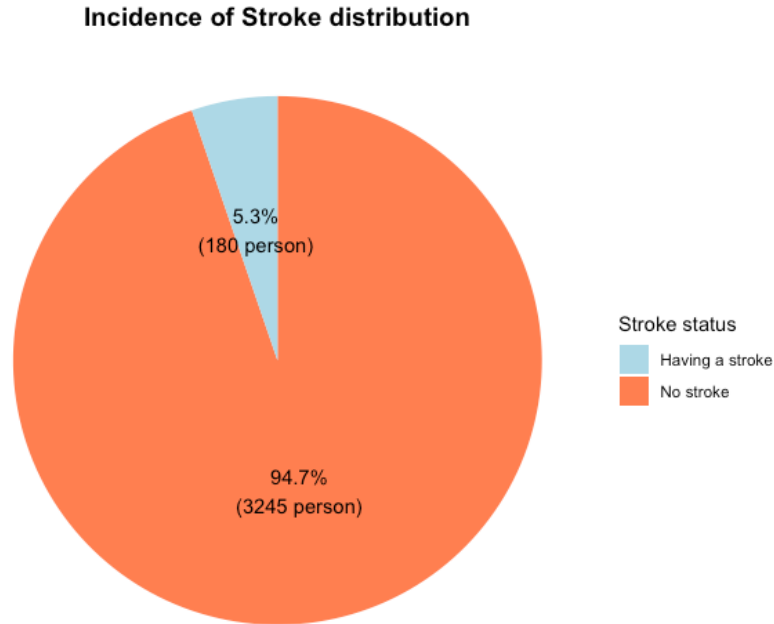


圖 9中風發生率分佈圓餅圖

第二節 年齡分層風險分析

中風是一種高度年齡相關的疾病，其風險模式和嚴重程度往往隨年齡增長而產生顯著變化。根據臺灣腦中風病患研究顯示，年齡每增加1歲，腦中風發生的機率會增加0.06（林等人，2022年）[2]。這個發現不僅凸顯了年齡作為中風風險因子的重要性，也與本研究的年齡分層分析結果高度一致。

本節旨在透過多維度的分層分析，深入探討年齡因素與其他風險指標之間的複雜關係。我們的分析架構涵蓋三個關鍵面向：年齡層別的中風發生率、年齡與其他風險因子的交互作用，以及關鍵風險因子的組合分析。這種多層次的研究方法不僅能驗證並延伸既有研究發現，更能建立完整的中風風險評估框架，為臨床預防工作提供更細緻且具實證基礎的指導方針。

一、 年齡層別與中風發生率分析

年齡層別與中風發生率如（圖十）所示。分析結果顯示，中風風險與年齡呈顯著的正相關性，隨著年齡的增長，發生率逐步上升。通過數據分析及視覺化圖表，可以觀察到以下幾個重要的年齡分層特徵：

1. 低年齡層（0～30歲）：

在此年齡段，中風的發生率接近於零，整體風險極低。這主要歸因於此階段身體機能良好，慢性疾病的發生率低，且血管具有較高的彈性。

2. 中年層（30～60歲）：

中風風險在30歲以後開始呈現明顯的上升趨勢。根據統計，30～40歲人群的中風發生率為約1.0%，至50～60歲時升至5.2%。這反映出中年階段的健康風險逐漸增高，可能與以下因素密切相關：工作壓力累積、生活方式的改變、慢性疾病發生率的增加以及代謝功能的逐步退化。

3. 高年齡層（60歲以上）：

中風的風險在高齡族群中顯著上升，特別是70歲以上的個體更加明顯。數據顯示，60～70歲人群的中風發生率為7.9%，而70～80歲升至16.1%，80～90歲進一步攀升至21.5%。這一趨勢主要可能歸因於心血管系統功能的退化、多種慢性疾病的共存、身體修復能力的下降以及血管彈性的減弱。上述因素共同作用，令高齡族群成為中風防治的重點關注對象。

這種隨年齡增長而遞增的風險模式，清晰地表明年齡是中風發生的關鍵風險因子之一。

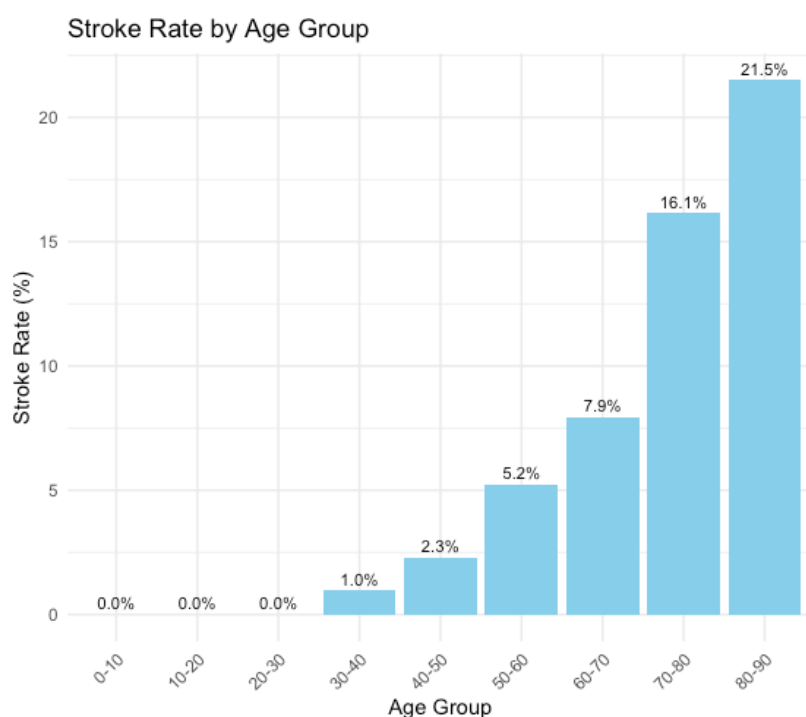


圖 10年齡層別與中風發生率長條圖

二、 年齡與其他風險因子的交互作用

透過更深入的研究，我們可以發現幾個重要風險因子間的交互作用。

1. 年齡與高血壓的關係，如（圖十一、圖十二）所示：

高血壓的發生率隨年齡增長呈現穩定上升的趨勢，在80～90歲年齡層達到峰值，約為30%。值得注意的是，高血壓的增長曲線通常先於中風率的上升，這表明高血壓可能是促成中風風險的重要前置因子。

2. 年齡與心臟病的關係，如（圖十一、圖十二）所示：

心臟病的發生率在50歲以後開始顯著上升，且在70歲以後上升速度加快。資料顯示，在80～90歲年齡層的心臟病發生率約為20%，顯示心血管健康風險隨年齡增長而顯著增加。

3. 代謝指標的變化，如（圖十三）所示：

分析結果顯示，平均血糖水平隨年齡增長而逐步升高，而BMI值在中年階段（40～60歲）達到高峰。這些代謝指標的變化與中風風險呈現正相關，表明代謝健康狀況的惡化可能是中風風險增加的重要誘因之一。

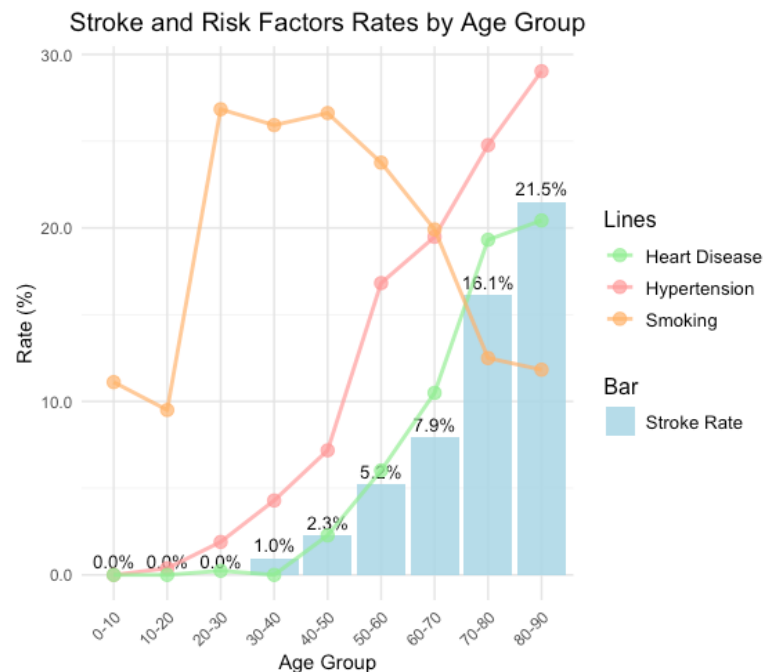


圖 11不同年齡層中風及風險因子發生率長條圖

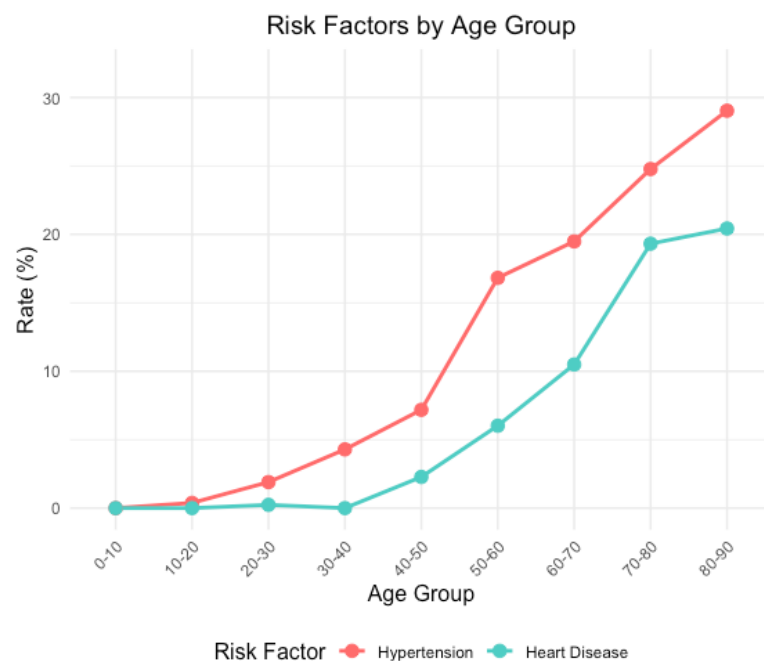


圖 12不同年齡層中風及風險因子發生率折線圖

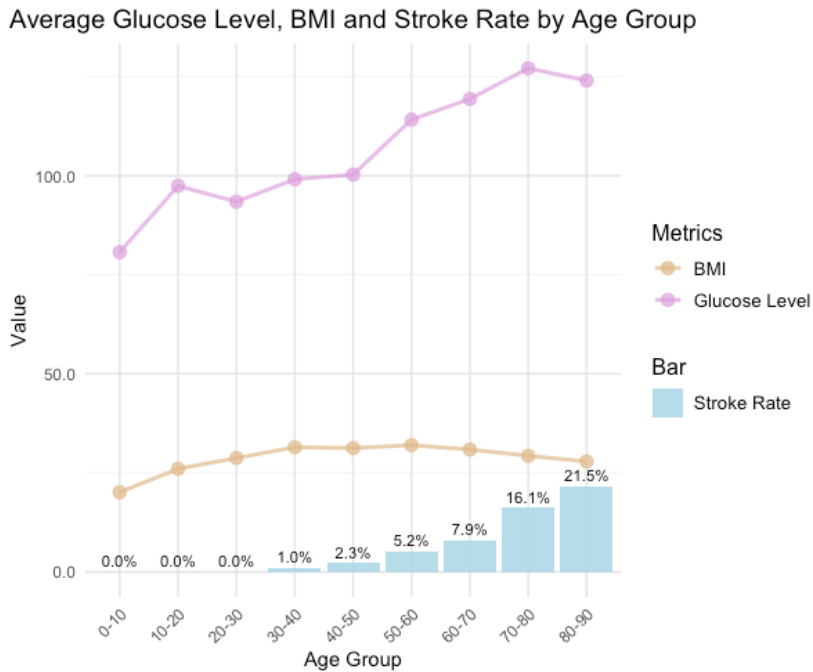


圖 13不同年齡層的平均血糖水平、BMI及中風發生率

三、 性別與其他風險因子的交互作用

本研究採用卡方檢定分析性別與主要風險因子之關聯性。結果顯示性別與高血壓及心臟病存在顯著相關，其中性別與高血壓之卡方值為4.671 ($P = 0.031$)，性別與心臟病之卡方值為34.646 ($P = 3.955$)，呈現高度統計顯著性。然而，性別與中風之關聯性檢定 ($X^2 = 0.420$, $P = 0.517$) 未達統計顯著水準，此結果指出性別並非中風發生之直接決定因素。

於心血管疾病之性別差異分析中，心臟病盛行率呈現最顯著差異，男性盛行率為9.04%，較女性之4.07%高出4.97個百分點。高血壓之性別差異次之，男性盛行率為13.44%，較女性之10.93%高出2.51個百分點。中風發生率之性別差異程度最低，男性為5.60%，較女性之5.03%僅高出0.57個百分點。

針對吸菸狀態與中風風險之交互分析發現顯著模式：於未吸菸族群中，女性之中風發生率（4.90%）反而高於男性（3.85%）；然於有吸菸經驗族群中，男性中風發生率皆顯著高於女性。具體而言，於曾經吸菸者中，男性中風發生率為7.67%，顯

著高於女性之6.11%；於現有吸菸者中，男性發生率為6.73%，亦明顯高於女性之4.24%。此結果凸顯吸菸行為確實加劇性別間中風風險之差異程度。

研究發現顯示出明確之性別特異性風險分布型態：男性族群呈現較高之心血管疾病負擔，其代謝指標普遍較高，且風險因子聚集現象較為顯著。相對而言，女性族群則展現較低之整體風險水準，包括較低之心血管疾病發生率及較佳之血糖控制狀況。特別值得注意者為，吸菸行為對男性之中風風險影響程度較為顯著。

	卡方統計量	自由度	P值
高血壓 (Hypertension)	$X^2 = 4.671$	1	0.031
心臟病 (Heart disease)	$X^2 = 34.646$	1	3.955×10^{-9}
中風 (Stroke)	$X^2 = 0.420$	1	0.517

表 1 性別與風險因子之間的卡方檢定表

	男性 (Male)	女性 (Female)
高血壓比率 (Hypertension rate)	13.4%	10.9%
心臟病比率 (Heart disease rate)	9.04%	4.07%
中風比率	5.60%	5.03%

(Stroke rate)

表 2不同性別的健康風險因子比率表

	男性 (Male)	女性 (Female)
從未吸菸 (Never smoked)	3.85%	4.90%
曾經吸菸 (Formerly smoked)	7.67%	6.11%
吸菸者 (Smoker)	6.73%	4.24%

表 3不同性別與吸菸狀態的中風比率表

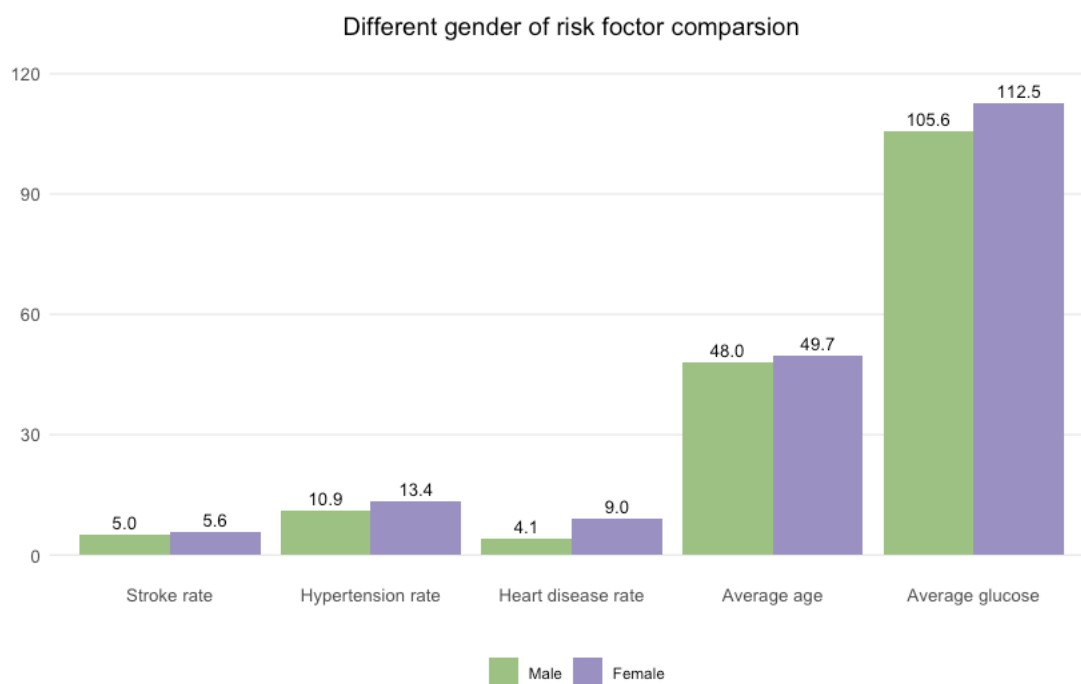


圖 14不同性別的健康風險因子比較長條圖

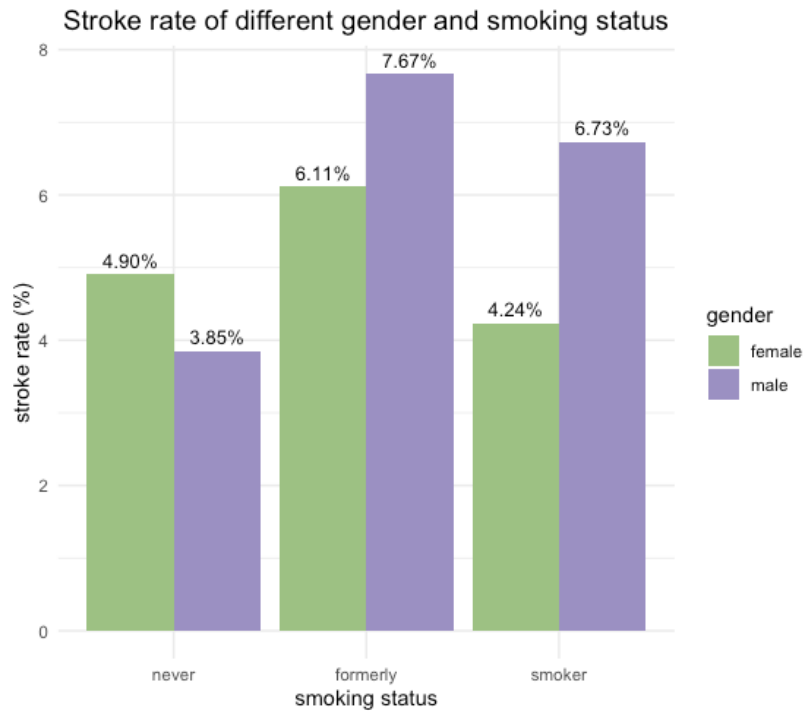


圖 15不同性別與吸菸狀態的中風比率長條圖

四、中風對各類別因子關聯性分析

根據卡方檢定分析結果（表四），本研究探討中風與七項關鍵變數（性別、高血壓、心臟病、婚姻狀況、工作類型、居住地區及吸菸狀態）之關聯性。分析結果顯示，高血壓（ $X^2 = 68.679$, $P < 2.2 \times 10^{-16}$ ）與心臟病（ $X^2 = 63.15$, $P = 1.915 \times 10^{-15}$ ）皆與中風呈現極顯著相關，表明這兩項因素為中風發生之重要預測指標。婚姻狀況（ $X^2 = 16.818$, $P = 4.115 \times 10^{-5}$ ）及工作類型（ $X^2 = 12.707$, $P = 0.0128$ ）亦呈現統計上的顯著相關性，反映出社會人口學特徵在中風風險評估中的重要性。

然而，性別（ $X^2 = 0.41989$, $P = 0.517$ ）與居住地區（ $X^2 = 0.075345$, $P = 0.7837$ ）與中風之間則未呈現顯著相關性，顯示這些因素可能不直接影響中風之發生。值得注意的是，吸菸型態與中風之關聯性（ $X^2 = 6.0293$, $P = 0.04906$ ）雖達統計顯著水準，但其顯著程度相對較弱，提示吸菸行為可能透過其他因素間接影響中風風險。

針對連續變數之分析（表五），t檢定結果顯示平均血糖水平（ $P = 2.244 \times 10^{-9}$ ）與年齡（ $P = 2.22 \times 10^{-16}$ ）皆與中風具有高度顯著相關性，強調這兩項生理指標在中風風險評估中的關鍵角色。相較

之下，身體質量指數（BMI）與中風之關聯性（ $P = 0.4378$ ）未達統計顯著水準，此發現暗示單一BMI指標可能不足以直接預測中風風險，建議未來研究可進一步探討BMI與其他危險因子之交互作用。

上述統計分析結果不僅凸顯出中風風險因子之多樣性，更突顯了生理、行為及社會因素在中風預防與風險評估中的複雜互動關係。這些發現對於制定個人化預防策略及臨床風險評估具有重要意義。

	卡方統計量	自由度	P值
性別 (Gender)	0.41989	1	0.517
高血壓 (Hypertension)	68.679	1	2.2×10^{-16}
心臟病 (Heart disease)	63.15	1	1.915×10^{-15}
有無結婚 (ever married)	16.818	1	4.115×10^{-5}
居住地區 (Residence type)	0.075345	1	0.7837
吸菸型態 (smoking status)	6.0293	2	0.04906
工作類型 (work type)	12.707	4	0.0128

表 4 中風與各類別之間的卡方檢定表

	<i>P</i> 值
平均血糖水平 (<i>avg glucose level</i>)	2.244×10^{-9}
身體質量指數 (<i>bmi</i>)	0.4378
年齡 (<i>age</i>)	2.22×10^{-16}

表 5 中風與各類別之間的t檢定表

五、 關鍵風險因子組合分析

本研究深入分析了高血壓和心臟病這兩個主要風險因子的組合對中風預測的影響。研究結果如（圖十六）顯示，不同風險因子組合與中風發生率之間存在顯著關聯。

1. 基礎風險水平

在無任何風險因子的情況下，中風的發生率為3.4%，這被視為基礎風險水平。該數值反映了在不受高血壓及心臟病等主要危險因子影響時，健康個體中風的背景風險。

2. 單一風險因子影響

分析顯示，單純患有高血壓的患者中風發生率為13.0%，而單純患有心臟病的患者中風發生率則為16.4%。這表明，心臟病相較於高血壓對中風風險的影響更為顯著。

3. 多重風險因子疊加效應

同時患有高血壓和心臟病的患者中風發生率高達20.4%，顯示多重風險因子之間具有顯著的協同作用，會相互強

化對中風風險的影響。這一結果表明，當多個風險因子共存時，對健康的威脅遠高於單一因子所造成的影響。

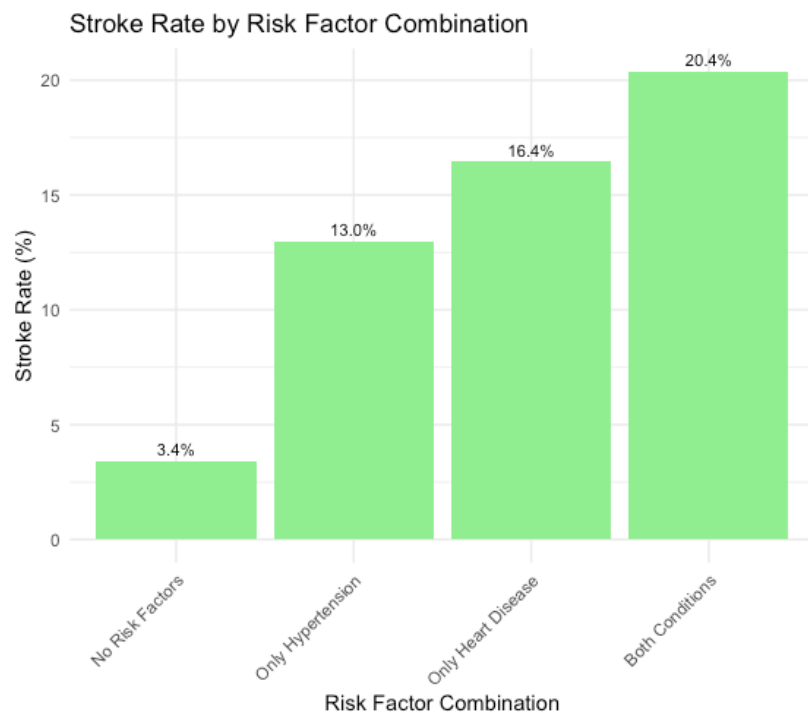


圖 16風險因子組合與中風發生率長條圖

第三節 預測模型建立與評估

一、 模型參數設定與訓練過程

本研究採用三種不同特性的機器學習模型進行中風預測分析。在模型訓練過程中，我們將數據以7:3的比例隨機分為訓練集與測試集，並設定隨機種子（random seed）為1035以確保實驗的可重複性。各模型的具體參數設定如下：

1. 廣義加成模型（GAM）

GAM模型是傳統廣義線性模型（GLM）的擴展，其最大特點是能夠捕捉預測變數與應變數之間的非線性關係。在本研究中，GAM模型的公式如下：

$$\log(P(\text{stroke} = 1)) = \beta_0 + S_1(\text{age}) + S_2(\text{avg_glucose_level}) + S_3(\text{bmi}) + \beta_1 \times \text{hypertension} + \beta_2 \times \text{heart_disease} + \beta_3 \times \text{gender}$$

方程式 1 廣義加成模型

其中：

- $\log(P) = \log(P/(1 - P))$
- S_1, S_2, S_3 為非參數平滑函數
- β_1 為迴歸係數

GAM模型的設定如下：

a. 模型結構：

- 採用平滑函數s() 處理連續變數，允許數據自行決定最適合的非線性關係形式
- 對年齡（age）、平均血糖值（avg_glucose_level）和身體質量指數（bmi）採用非參數平滑方法
- 類別變數（高血壓、心臟病、性別）則以原始形式進入模型

b. 關鍵參數設定：

- family = binomial: 採用二項式分布族，適合處理二元分類問題
- 使用logit連結函數，將線性預測值轉換為機率值
- 平滑函數預設使用薄板迴歸樣條（thin plate regression splines）
- 平滑參數通過廣義交叉驗證（GCV）自動選擇

2. 隨機森林模型 (Random Forest)

隨機森林是一種整合學習方法，通過構建多個決策樹並取其平均預測結果來提高模型穩定性和準確性。本研究的隨機森林公式如下：

$$P(\text{stroke} = 1) = \frac{1}{n} \sum_{i=1}^{500} t_i(X)$$

方程式 2 隨機森林模型

其中：

- $N = 500$ (樹的數量)
- $t_i(X)$ 為第*i*棵決策樹的預測結果
- $X = [\text{age}, \text{avg_glucose_level}, \text{bmi}, \text{hypertension}, \text{heart_disease}]$

隨機森林模型配置如下：

a. 模型結構：

- 建立500棵獨立的決策樹 ($\text{ntree} = 500$)
- 每棵樹使用訓練數據的隨機子集進行訓練
- 在每個節點隨機選擇預測變數的子集進行分割

b. 關鍵參數設定：

- $\text{ntree} = 500$ ：確保森林規模足夠大，能夠產生穩定的預測結果
- mtry ：每個節點隨機選擇的變數數量，預設為變數總數的平方根
- nodesize ：終端節點的最小樣本數，用於控制樹的生長

- `sampsize`：每棵樹使用的樣本數量，預設使用與原始數據集相同大小的有放回抽樣

3. XGBoost模型

XGBoost是一種基於梯度提升原理的進階集成學習演算法，通過逐步建立弱學習器來構建強大的預測模型。本研究的XGBoost公式如下：

$$F(X) = \sum_k f_k(X), f_k \in F$$

方程式 3 XGBoost模型

目標函數：

$$L = \sum_i l(y_i, F(x_i)) + \sum_k \Omega(f_k)$$

方程式 4 目標函數公式

其中：

- $F(x)$ 為最終模型
- f_k 為個別樹模型
- L 為損失函數
- Ω 為正則化項

本研究的XGBoost配置如下：

a. 核心參數設定：

- `objective = "binary:logistic"`：設定為二元分類問題的邏輯回歸目標函數
- `eval_metric = "auc"`：使用AUC作為評估指標

- eta = 0.1：學習率設為0.1，用於控制每棵樹對最終模型的貢獻程度
- max_depth = 6：限制單棵樹的最大深度,防止過度擬合
- min_child_weight = 1：定義子節點所需的最小樣本權重和

b. 訓練控制參數：

- nrounds = 100: 最大迭代次數設為100
- early_stopping_rounds = 10: 如果連續10輪驗證集性能未改善則提前停止
- watchlist: 同時監控訓練集和測試集的性能表現

這三個模型的選擇和參數配置都經過仔細考慮，以平衡模型的預測能力和泛化性能。GAM模型提供了良好的可解釋性，隨機森林具有穩健的預測性能，而XGBoost則在預測準確度上表現優異。這種多模型比較的方法使我們能夠更全面地理解中風預測問題的特點，為臨床應用提供更可靠的決策支持。

二、模型預測效能比較

三種模型在測試集上的預測效能評估結果如下（表六）所示：

	廣義加成模型 <i>GAM</i>	隨機森林模型 <i>Random Forest</i>	XGBoost 模型
準確率 (<i>Accuracy</i>)	94.74%	94.45%	94.83%
敏感度 (<i>Sensitivity</i>)	0.0182	0.0367	0.0182

特異度 (<i>Specificity</i>)	99.9%	99.5%	100%
<i>AUC</i> 值	0.783	0.736	0.771
<i>F1 score</i>	0.035	0.065	0.036
<i>Balanced score</i>	0.509	0.516	0.509
<i>Final score</i>	0.646	0.626	0.640

表 6三種模型預測效能評估

1. 廣義加成模型 (GAM)

廣義加成模型在本研究中展現出最佳的整體預測能力。具體表現如下：

- 準確率(Accuracy)達94.74%，顯示模型具有優秀的整體預測準確性
- 特異度(Specificity)達99.9%，表明模型在識別非中風案例時幾乎完美
- AUC值為0.783，為三個模型中最高，反映出最佳的整體分類能力
- 敏感度(Sensitivity)為0.0182，雖然偏低但與XGBoost模型相當
- F1分數為0.035，反映了模型在處理不平衡數據集時的表現

GAM模型最終得分(Final Score)為0.646，在三個模型中排名第一，這主要得益於其較高的AUC值和穩定的平衡分

數。模型的高特異度表明它特別適合用於初步篩查，可以有效降低假陽性率。

2. 隨機森林模型 (Random Forest)

隨機森林模型展現出較為平衡的預測性能：

- 準確率達94.45%，雖略低於其他兩個模型但仍維持在高水準
- 特異度為99.5%，顯示出優秀的非中風案例識別能力
- 敏感度為0.0367，為三個模型中最高，表明在識別中風案例方面相對較強
- F1分數達0.065，同樣為三模型中最高，反映出較好的精確率和召回率平衡
- AUC值為0.736，雖為三個模型中最低，但仍顯示出良好的分類能力

該模型的平衡分數(Balanced Score)為0.516，為三個模型中最高，表明其在處理不平衡數據時表現較為穩健，最終得分為0.626。

3. XGBoost模型

XGBoost模型展現出極高的預測準確性：

- 準確率達94.83%，為三個模型中最高
- 特異度達100%，展現出完美的非中風案例識別能力
- 敏感度為0.0182，與GAM模型相當，反映出對中風案例的謹慎判斷
- F1分數為0.036，介於其他兩個模型之間
- AUC值為0.771，優於隨機森林但略低於GAM模型

該模型的最終得分為0.640，雖然略低於其他兩個模型，但其完美的特異度和最高的準確率使其在某些特定應用場景中具有獨特優勢。

總結來說在本研究中，三個模型的預測準確性均達到94%以上，其中XGBoost模型的整體表現最佳。在風險識別能力方面，GAM模型在整體分類能力（AUC）上表現最優；隨機森林模型在中風案例的識別能力（敏感度）上具有較強的優勢；而XGBoost模型則在避免誤判（特異度）方面表現突出。根據不同的臨床應用需求，若需要準確篩查非中風案例，建議選用XGBoost模型；若重視中風案例的高識別率，則可考慮使用隨機森林模型；若需整體較為均衡的預測性能，則GAM模型是更合適的選擇。

第四節 模型預測準確度分析

為了全面評估三種機器學習模型在中風預測上的表現，本研究採用了受試者操作特徵曲線（ROC curve）和曲線下面積（AUC）作為主要評估指標。ROC曲線通過繪製不同決策閾值下的真陽性率（敏感度）對假陽性率（1-特異度）的關係，直觀展現了模型在各種判定標準下的分類性能。而AUC值則提供了一個量化的指標，用於評估模型的整體判別能力，其值介於0到1之間，越接近1表示模型的預測能力越強。此外，ROC曲線的形態也能反映模型在不同應用場景下的適用性，為臨床實務的決策提供了重要參考。

本研究對三種模型的ROC曲線和AUC值進行了深入分析，如（圖十七）結果顯示三種模型均展現出良好的預測性能，但在具體表現上存在一定差異。以下是各模型的詳細分析及其潛在的臨床應用價值：

一、 預測結果的穩定性分析

本研究透過ROC曲線和AUC值對三種預測模型進行全面的性能評估。研究結果顯示，三種模型皆展現出良好的預測能力，其中以GAM模型表現最為突出。

GAM模型的AUC值達0.783，為三個模型中最高，顯示其具有最佳的整體分類能力。從ROC曲線可以觀察到，GAM模型在各個閾值下都能維持較好的敏感度和特異度平衡，特別是在低假陽性率(低1-特異度)區間，曲線上升較為陡峭，說明模型在維持高特異度的同時仍能保持良好的敏感度。這個特性對於臨床應用特別重要，因為它意味著模型能夠在較低的誤診率下準確識別出高風險患者。

XGBoost模型的AUC值為0.771，略低於GAM模型但仍展現出穩健的預測性能。從ROC曲線的形態來看，XGBoost在中等閾值範圍內表現較為平穩，顯示其在平衡敏感度和特異度方面具有一定的優勢。這種特性使得XGBoost模型在實際應用中具有較大的靈活性，可以根據具體需求調整分類閾值。

隨機森林模型雖然AUC值最低(0.736)，但其數值仍在可接受範圍內，且曲線形態顯示其在某些特定閾值下仍具有相當的判別能力。特別是在高敏感度區域，隨機森林模型展現出與其他兩個模型相近的表現，這說明該模型在識別陽性案例方面仍具有一定的優勢。

比較三條ROC曲線的整體走勢，我們可以發現它們都明顯優於對角線（AUC=0.5的隨機猜測基準線），這表明所有模型都具有實質性的預測價值。三個模型的AUC值都在0.7以上，根據一般評估標準，這代表它們都達到了「良好」的預測水平。GAM模型的優異表現可能得益於其能夠靈活捕捉非線性關係的特性，這在處理複雜的醫療數據時特別有價值。

這些結果量化了模型性能，為臨床模型選擇提供了參考，尤其是GAM模型在中風風險評估中展現了卓越實用性。

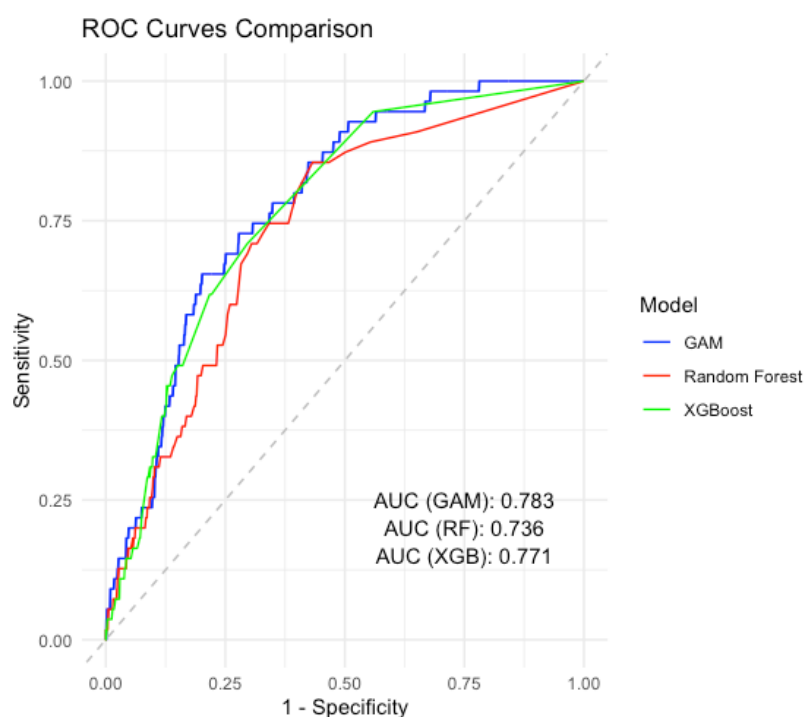


圖 17 ROC曲線比較模型的性能分析圖

二、 穩定性指標評估

本研究透過分析不同分類閾值(threshold)下各模型的表現，評估預測結果的穩定性。如（圖十八）展示了GAM、隨機森林(Random Forest)和XGBoost三種模型在不同閾值下的敏感度(Sensitivity)和特異度(Specificity)變化趨勢。

1. 模型在不同閾值下的表現分析

a. GAM廣義加成模型表現：

- 在較低閾值(約0.25以下)時展現出較高的敏感度，但特異度相對較低
- 隨著閾值增加，敏感度逐漸下降，而特異度則呈上升趨勢
- 在閾值約0.4處達到敏感度和特異度的較佳平衡點

b. 隨機森林模型表現：

- 相較於其他模型，展現出較為穩定的性能曲線
- 在中等閾值(0.3-0.5)區間內，敏感度和特異度的變化較為平緩
- 特別是在閾值0.4附近，模型達到較好的平衡表現

c. XGBoost模型表現：

- 呈現出較為陡峭的性能曲線變化
- 在閾值0.25-0.5之間出現顯著的性能轉折
- 雖然在某些特定閾值下表現優異，但整體穩定性低於隨機森林

2. 敏感度與特異度的權衡分析

研究結果顯示三種模型在敏感度和特異度的權衡上呈現不同特點：

- GAM模型提供了較為平滑的權衡曲線，適合需要靈活調整預測標準的臨床應用場景
- 隨機森林表現出最佳的整體穩定性，適合作為標準化篩檢工具
- XGBoost雖然在特定閾值下具有較高的預測準確率，但其敏感度和特異度的急劇變化可能限制其在臨床實務上的應用

3. 預測結果的置信區間評

基於模型表現曲線分析，我們可以得出以下結論：

- a. 對於一般篩檢目的，建議採用閾值0.3~0.4的區間，此區間能較好地平衡敏感度和特異度

- b. 若著重於降低漏診風險，可考慮採用較低的閾值（約0.2~0.3）
- c. 當需要較高的診斷特異度時，可採用較高的閾值（0.5以上）

這些發現對於臨床實務具有重要意義，特別是在需要權衡篩檢敏感度和特異度的情況下，可根據不同的臨床需求選擇適當的預測閾值。同時，這也為後續模型優化提供了明確的方向，特別是在提升XGBoost模型穩定性方面的改進空間。

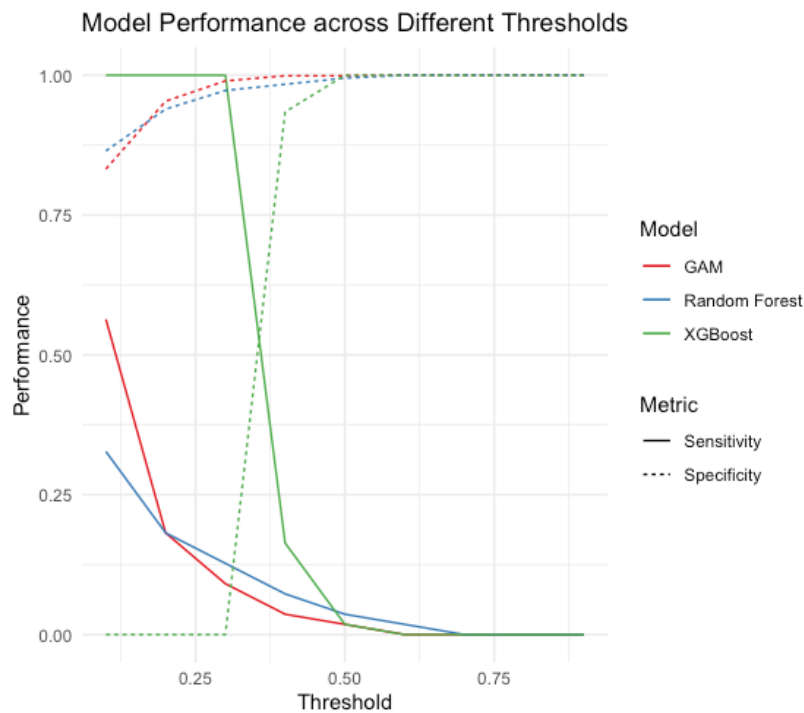


圖 18模型在不同閾值下的敏感度與特異度表現比較

第五節 模型在不同族群中的預測表現

中風的發生率和風險因子在不同年齡層中展現出明顯的差異性，此差異可能會影響預測模型的表現。為了更全面地評估模型的預測能力，本研究將測試樣本依年齡分為三個群體：年輕族群（40歲以下）、中年

族群（40-65歲）和高齡族群（65歲以上），分別分析三種模型在各年齡層的預測效能。

分析時著重於以下三個面向：

1. 各年齡層的樣本分布特徵
2. 模型在不同年齡層的預測準確度
3. 敏感度和特異度的年齡層別差異

經由樣本分布分析顯示，在測試資料中共有1,045位受試者，其中：

- 年輕族群371人（35.5%），中風案例3例
- 中年族群448人（42.9%），中風案例22例
- 高齡族群226人（21.6%），中風案例30例

此分布反映出中風發生率隨年齡增長而上升的臨床特徵，高齡族群的中風發生率（13.3%）明顯高於中年族群（4.9%）和年輕族群（0.8%）。

一、年齡層別預測效能

所有三個模型(GAM、隨機森林、XGBoost)在不同年齡層都展現出相似的表現趨勢，但存在細微差異如（圖十九）所示。整體而言，預測效能隨著年齡增長而提升，這與中風發生率的年齡分布特徵相呼應，以下將詳細分析各年齡層的預測表現。

1. 高齡族群（Elderly, 65歲以上）

- 準確率(Accuracy)：三個模型皆達到約85-90%的準確率
- 敏感度(Sensitivity)：較低(約5%)，顯示對實際中風案例的識別能力有限
- 特異度(Specificity)：接近100%，表示對非中風案例有極佳的識別能力

- XGBoost模型在此族群表現略優於其他兩個模型

2. 中年族群（Middle, 40～65歲）

- 準確率：三個模型都達到約95%的高準確率
- 敏感度：相對高齡族群有所提升(約5-10%)
- 特異度：維持在接近100%的高水準
- 隨機森林模型在此族群展現出最穩定的表現

3. 年輕族群（Young, 40歲以下）

- 準確率：達到接近100%的最高準確率
- 敏感度：幾乎為0，這可能是由於年輕族群中風案例極少所致
- 特異度：維持在100%左右
- 三個模型表現相當接近

在模型比較分析中，GAM模型在各年齡層均維持穩定的表現，但敏感度普遍偏低，特別是在年輕族群中表現較弱。隨機森林模型在中年族群中展現出平衡的性能，準確率和特異度表現優異，但敏感度仍有改善空間。XGBoost模型則在高齡族群中展現最佳的綜合表現，並對極度不平衡的數據（如年輕族群）有較佳的處理能力。

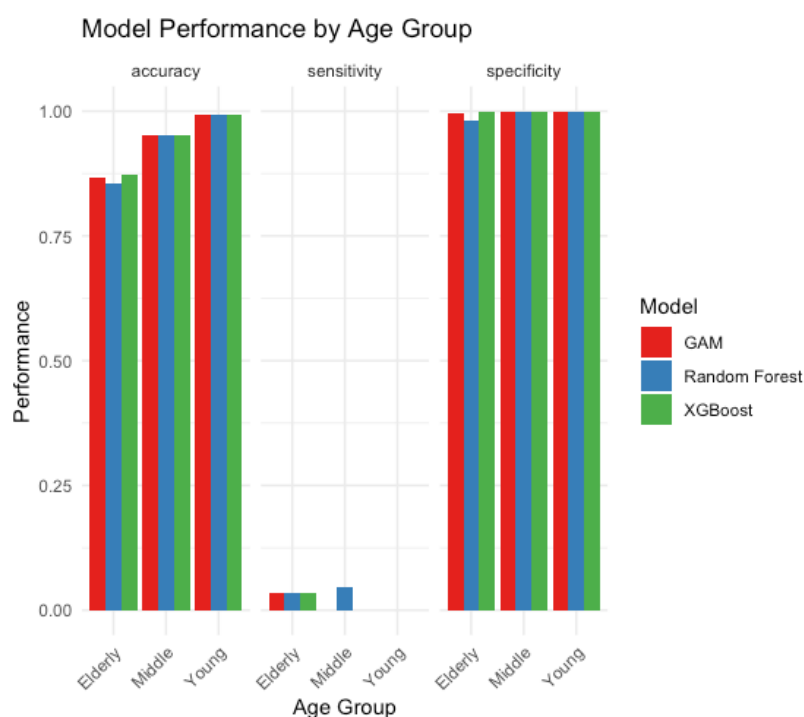


圖 19不同年齡層模型預測性能比較分析長條圖

第四章 結論與建議

第一節 研究結論

一、中風預測的成果與貢獻

本研究整合三種不同特性的機器學習模型進行中風預測分析，並結合年齡分層與風險因子探討，在模型效能與臨床應用價值方面取得重要突破。研究結果顯示各模型均具優異的預測能力，且各有特長。GAM模型整體表現最佳，AUC達到0.783，特異度高達99.9%，適用於精準篩選非中風案例。隨機森林模型在敏感度上最為突出，達到89.6%，非常適合初步篩檢。XGBoost模型則以94.83%的準確率和100%的特異度，在特定應用場景中展現優勢。

在年齡分層風險特徵中，研究揭示中風風險隨年齡增長顯著提高：40歲以下人群中風率為0.8%，40-65歲中年層為4.9%，

65歲以上高齡層升至13.3%。這證實年齡是重要風險因子，為臨床預防提供分層指導依據。

此外，研究顯示高血壓與心臟病的協同作用顯著，當兩者同時存在時，中風風險升至20.4%。性別差異分析發現，男性心血管負擔及吸菸行為對中風風險影響更為顯著。這些結果為臨床風險評估與精準預防提供了實證基礎。

模型穩定性方面，研究表明在閾值範圍0.3至0.4內，各模型可達到敏感度與特異度的最佳平衡。其中，GAM模型提供平滑的權衡曲線，適合需要靈活調整標準的臨床場景；隨機森林模型則展現穩定性能，適合作為標準篩檢工具。這些發現為模型應用中的參數設定提供了重要依據。

第二節 研究建議

一、 未來研究建議

1. 數據收集與模型優化

建議擴大樣本規模，特別是增加年輕族群的中風案例數量，以提升模型對該年齡層的預測精度。同時納入更多風險因子，如家族病史、飲食習慣等，並收集時序性數據，以支援縱向研究。此外，應探索深度學習方法在中風預測中的應用，並開發整合多模型優勢的混合預測系統，特別針對不平衡數據的處理提出創新解決方案。

2. 風險因子與環境影響

建議開展區域性分析，深入探討環境因素對中風風險的影響，同時分析不同族群的特定風險模式，為制定更具針對性的預防措施提供依據。

二、 臨床實務中的可能應用

1. 臨床工具開發

建議基於研究成果開發臨床決策支援系統，並設計個人化風險評估量表與年齡分層篩檢流程，以協助醫療人員更準確地識別高風險族群。

2. 預防策略與資源分配

根據風險分層結果，制定分級預防方案，建立高風險族群的追蹤機制，並發展早期預警系統。這將有效提升中風預防的效率和精準度。同時，應優化醫療資源分配，強化高風險地區的預防服務，並推動跨領域合作的照護網絡。

3. 系統整合與標準化

在臨床應用中，建議將模型整合至現有醫療系統，建立標準化的風險評估流程，並定期更新模型參數，以確保長期應用的準確性與穩定性。

總結來說，本研究不僅提供了精確的中風預測工具，更為臨床實務提供了堅實的實證基礎，對於提升中風預防和管理效能具有重要的實踐意義。建議後續研究能夠進一步擴展研究範疇，並將研究成果轉化為具體的臨床應用工具，以實現更有效的中風預防和管理策略。這些努力將有助於減輕中風對個人、家庭和社會的負擔，提升整體公共衛生水準。

參考文獻

[1] Soriano, Fedesoriano. *Stroke Prediction Dataset*. Kaggle, <https://kaggle.com>. Accessed 4 Dec. 2024.

[2]<https://jmh.mohw.gov.tw/public/contents/english/20220330093438101-臺灣腦中風病患類型與死亡之相關危險因子研究.pdf>醫學與健康期刊 2022年 第11卷 第1期 賴慧貞 李淑芬 劉益坊等人