



# 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

613890218 數據科學所碩一 施宇  
鴻613890069 數據科學所碩一 鄭  
傑丞



# CONTENTS

1. 研究背景與動機
  2. 研究目的
  3. 研究方法
  4. 研究結果
  5. 結論與建議
- 

01

## 研究背景與動機

# 研究背景

腦中風(stroke)是全球主要的死亡和失能原因之一，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。

# 中風問題的嚴重性



## 高致死率：

中風是全球主要的致死和失能原因之一，若未能及時處理，可能造成永久性損傷甚至導致生命危險。

## 高復發率：

一次中風後，患者再次中風的風險顯著增加，復發的中風往往會更嚴重，恢復難度更高。

## 台灣現況：

中風是台灣十大死因之一。根據統計，台灣每年約有6萬至7萬人發生中風，相當於每10分鐘就有1人中風。

# 年齡分層的必要性

## 年齡對風險影響：

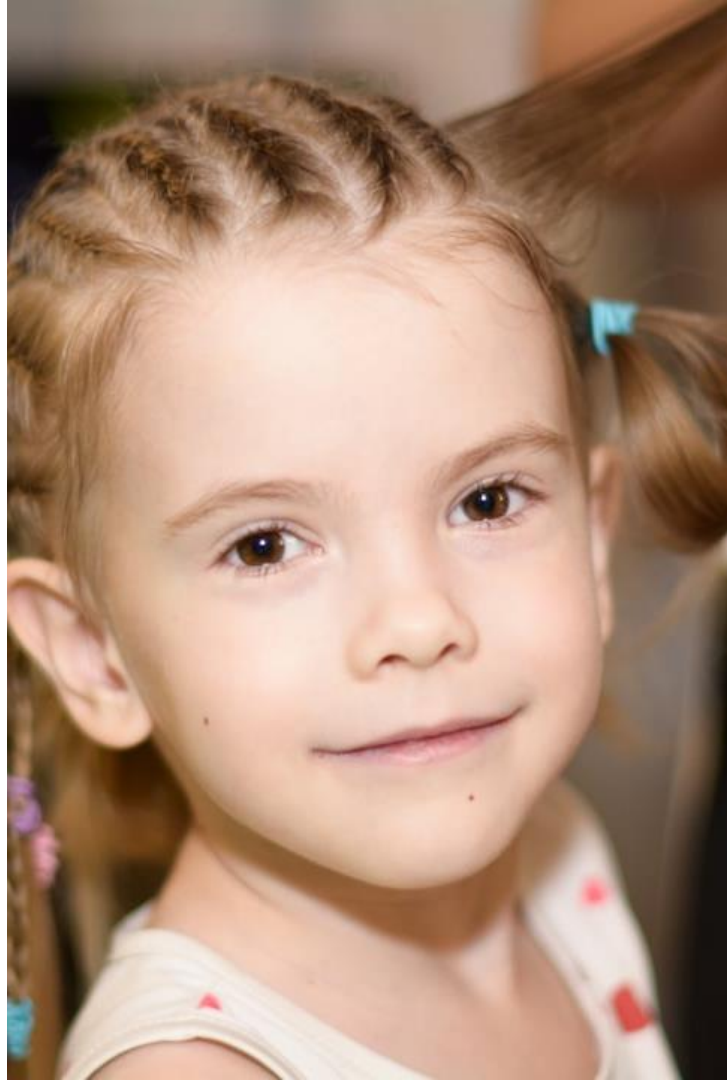
不同的年齡層的中風風險、發生率呈顯著差異

## 高危族群識別：

透過年齡分層分析，識別出易於中風的族群。

## 預測模型的分層表現：

透過不同年齡層的分析，可以提高模型預測準確度



# 風險因子



## 主要危險因子：

年齡、高血壓、心臟病、血糖濃度、BMI等風險因素密切相關。

## 風險因子交互作用：

透過分析找出增強中風風險的關鍵風險因子

## 風險評估框架：

建立全面的風險評估模型，考慮多種風險因子間的影響。

02

## 研究目的



# 研究目的

## 建立多元預測模型：

廣義可加模型(GAM)、隨機森林(Random Forest)及極限梯度提升(XGBoost)演算法。

## 評估關鍵的風險因子：

1. 平均血糖濃度(Average Glucose Level)
2. 高血壓(Hypertension)
3. 心臟病(Heart Disease)
4. 年齡(Age)

## 模型效能評估與比較：

建立受試者操作特徵(ROC)曲線

計算曲線下面積(AUC)

分析各模型的敏感度(Sensitivity)和特異度(Specificity)

# 評估關鍵風險因子



我們特別著重分析四個主要預測變數：

- 平均血糖濃度
- 高血壓
- 心臟病
- 年齡

研究探討上述變數與中風風險間的影響程度及關聯性

# 模型建立目標

## 應用統計模型：

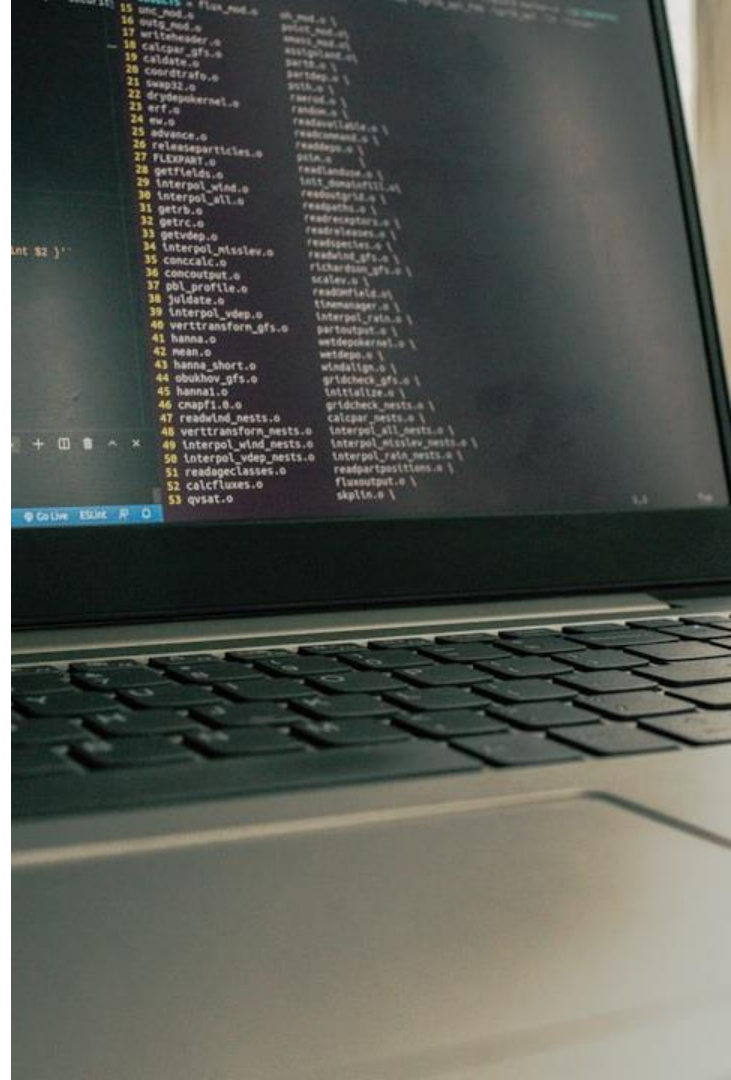
使用廣義線性加成(GAM)、隨機森林和XGBoost  
建立預測模型。

## 年齡與風險因子分析：

針對不同年齡層的風險因子進行深入探討。

## 模型效能比較：

評估三種模型在中風預測上的準確性和效能。



# 效能評估方法

## ROC曲線：

建立並分析ROC曲線。

## AUC計算：

計算並比較不同模型的AUC值。

## 敏感度與特異度：

檢驗各模型在預測中的敏感度和特異度。



03

## 研究方法

# 研究方法

01

**數據來源與處理：**

確保資料的結構和有效性。

02

**預測模型設計：**

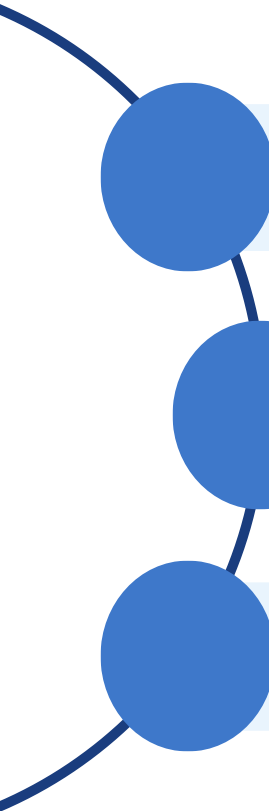
設定各模型參數。

03

**模型選擇：**

選擇最佳預測模型方法。

# 數據來源與處理



## 資料集來源:

使用Fedesoriano在Kaggle上的中風資料集進行分析。

## 資料預處理:

刪除無意義的欄位並處理缺失值

## 最終觀測值:

確保只有完整的數值進入模型分析，共計3,425筆資料。

# 預測模型設計

## 資料分割策略：

將資料以7:3的比例分割為訓練集與測試集。

## 設定隨機種子：

確保實驗的可重複性，隨機種子設定為1035。

## 模型參數設計：

各模型設定學習率、最大樹深度等重要參數。





# 模型選擇

## GAM的選擇：

此模型適合連續變數的平滑處理，為傳統統計學的代表。

## 隨機森林的特點：

樹狀結構能夠捕捉複雜的交互作用。

## XGBoost的優勢：

針對大資料集的高效能和準確性，表現出色。



04

## 研究結果

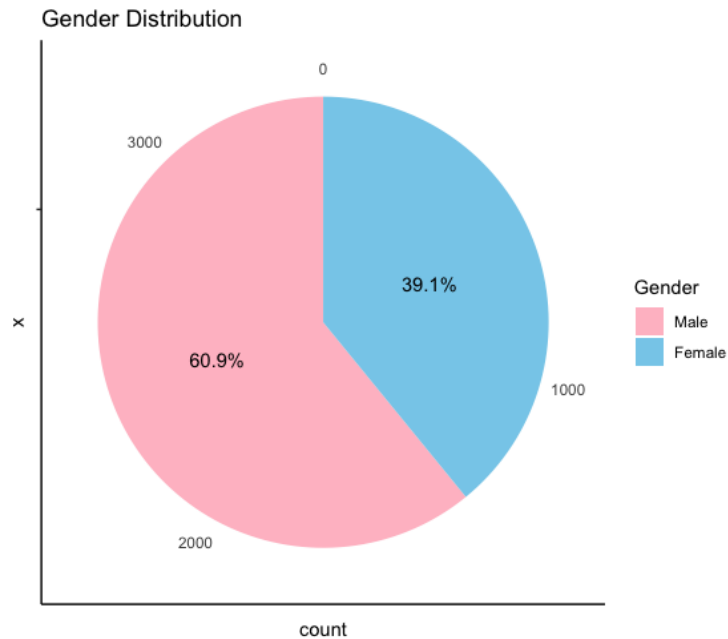
04-  
1

## 敘述統計

# 性別分布

性別分布顯示如圖所示，女性樣本共有2,086人（60.9%），男性樣本共有1,339人（39.1%）。

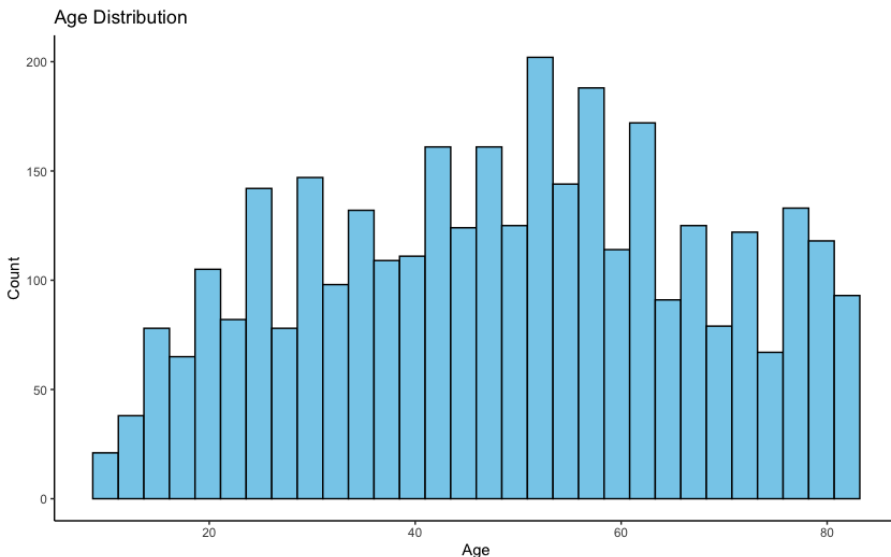
女性比例略高於男性，但整體性別分布仍保持在合理範圍內。



# 年齡分布

年齡分布如圖所示涵蓋從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。

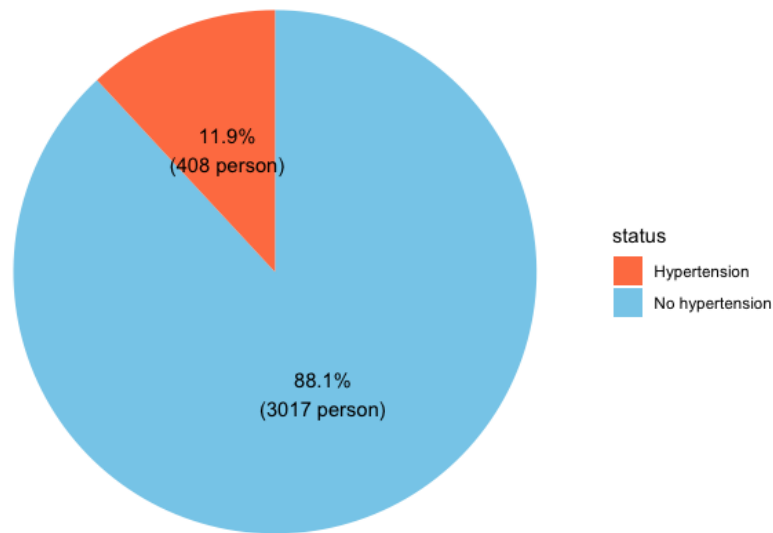


# 高血壓狀況

如圖所示，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。

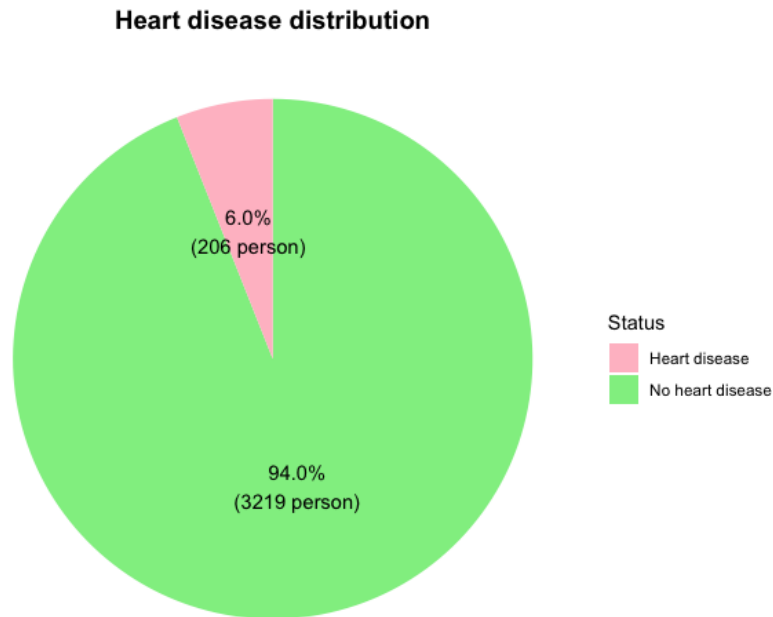
Hypertension status distribution



# 心臟病史

如圖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%)。

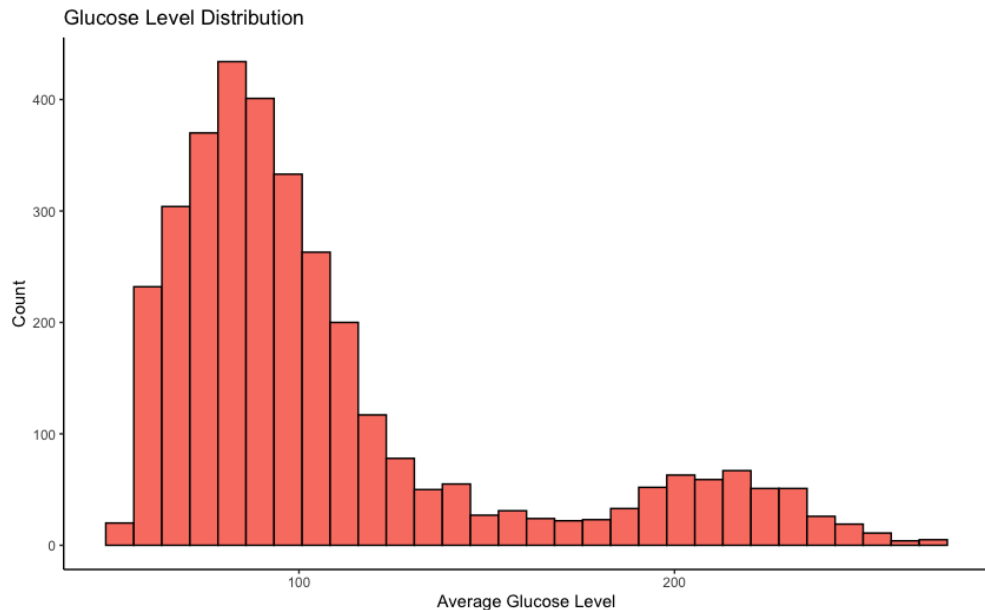
資料顯示絕大多數樣本無心臟病，僅有少部分  
樣本為心臟病患者。



# 平均血糖濃度

平均血糖濃度的分佈情形如圖所示，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。

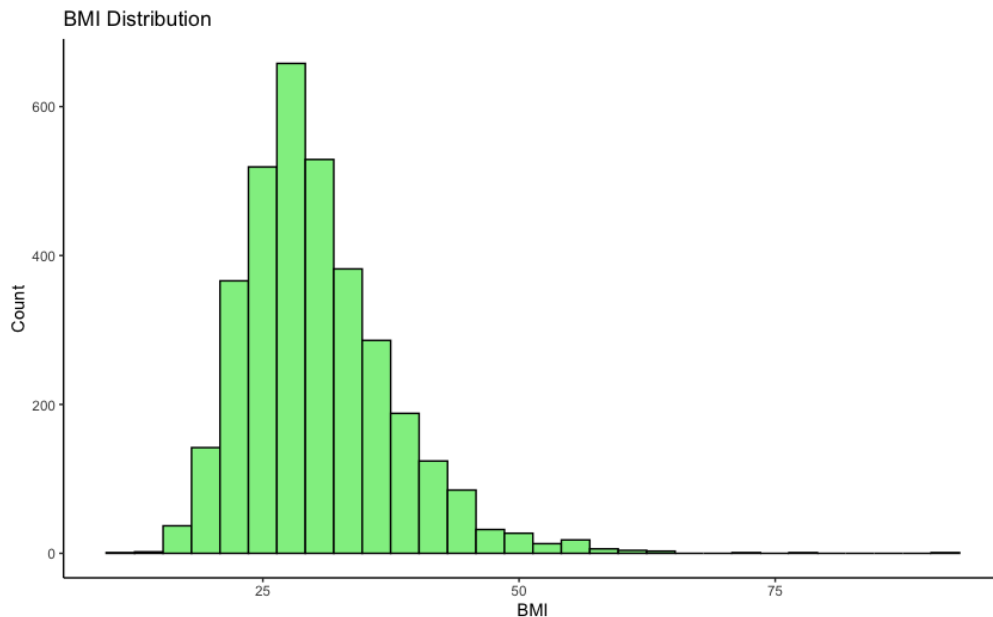




# 身體質量指數 (BMI)

身體質量指數的分佈情形如圖所示，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。

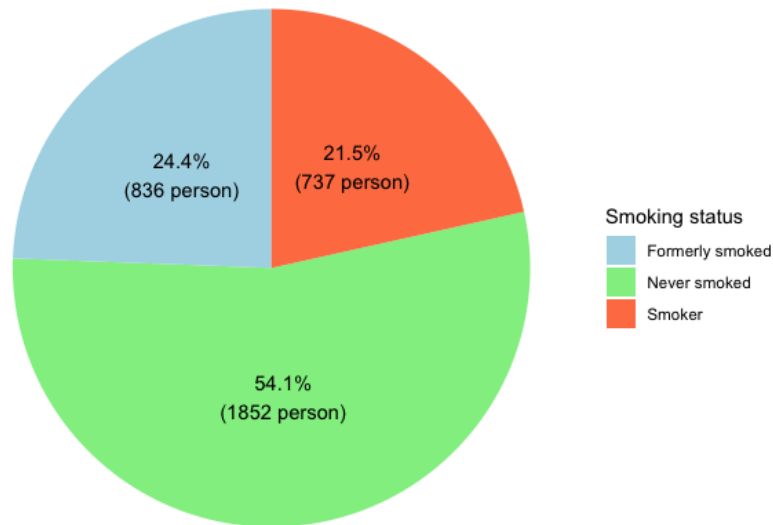


# 吸菸狀況

吸菸狀況如圖所示，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。

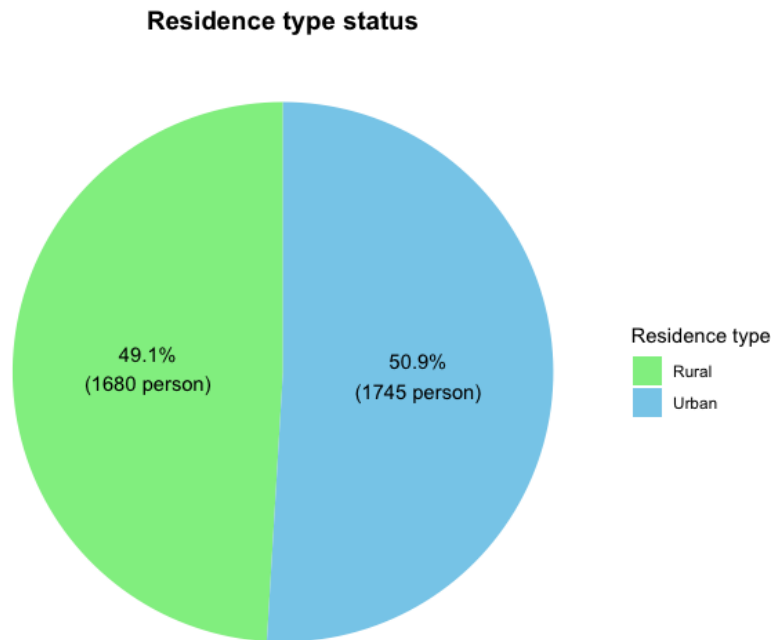
Smoking status distribution



# 居住類型

樣本中居住類型如圖所示，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。

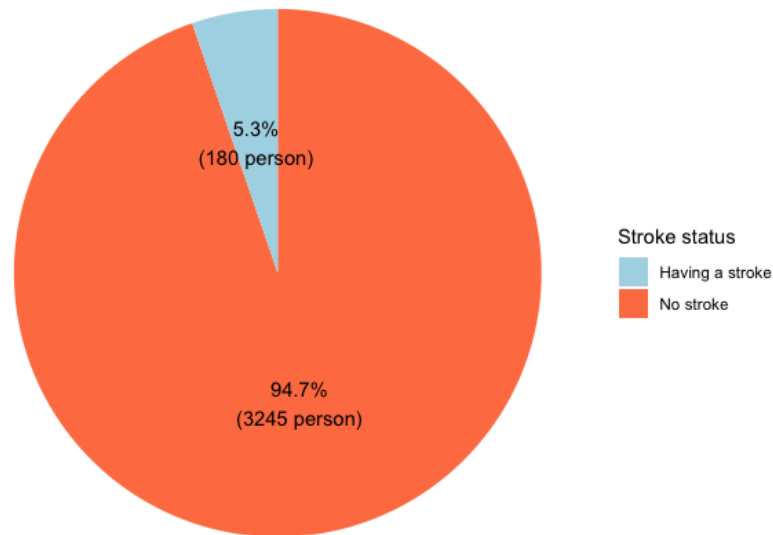


# 中風發生率

整體樣本的中風發生率分佈如圖所示，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。

Incidence of Stroke distribution



04-2

## 年齡分層風險分析

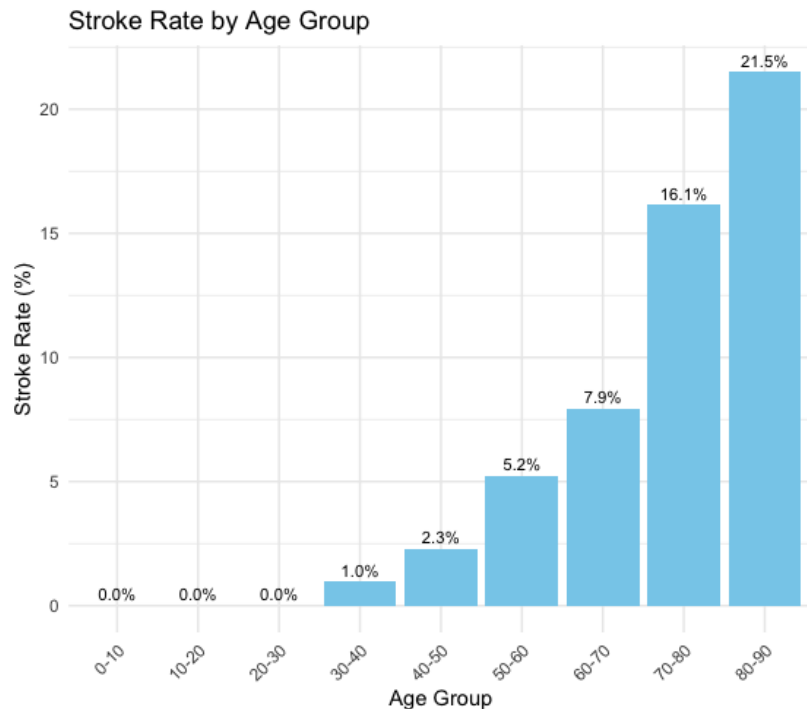
# 年齡層別與中風發生率分析

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。

低年齡層（0～30歲）：中風的發生率接近於零，整體風險低。

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。這反映出中年階段的健康風險逐漸增高。

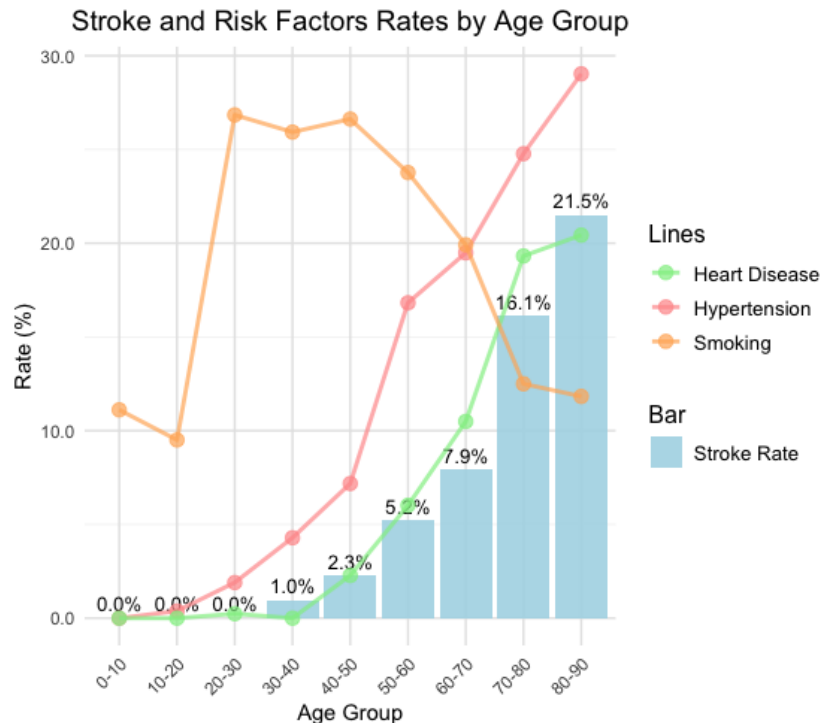
高年齡層（60歲以上）：中風的風險在高齡族群中顯著上升，特別是70歲以上的個體更加明顯。



# 不同年齡層中風及風險因子發生率長條圖

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%。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。

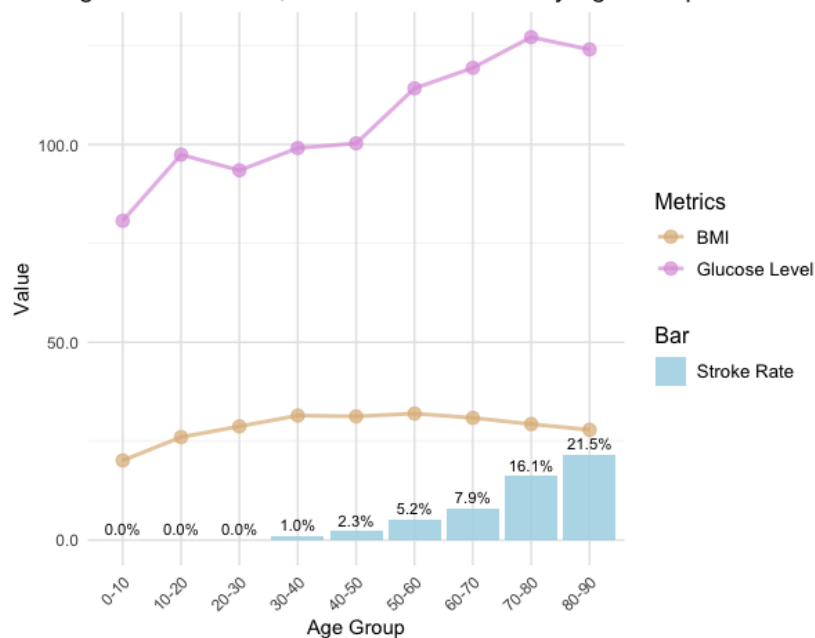
心臟病的發生率在50歲以後開始顯著上升，且在70歲以後上升速度加快。資料顯示，在80～90歲年齡層的心臟病發生率約為20%，顯示心血管健康風險隨年齡增長而顯著增加。



# 代謝指標的變化

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。

Average Glucose Level, BMI and Stroke Rate by Age Group





# 性別與其他風險因子的交互作用

## 整體風險因子比較 (男性 VS 女性)

中風發生率: 5.60% VS 5.03%

高血壓比率: 13.44% VS 10.93%

心臟病比率: 9.04% VS 4.07%

平均血糖值: 112.49 VS 105.63 MG/DL

## 吸菸狀態與性別的交互作用

從未吸菸: 男性反而較低 (3.85% VS 4.90%)

曾經吸菸: 男性明顯較高 (7.67% VS 6.11%)

現有吸菸: 男性顯著較高 (6.73% VS 4.24%)

男性族群: 心血管疾病負擔較重, 代謝指標較高

女性族群: 整體風險較低, 血糖控制較佳

特別發現: 吸菸對男性中風風險的影響更為顯著

# 性別與其他風險因子的交互作用

	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 (Stroke rate)	5.60%	5.03%

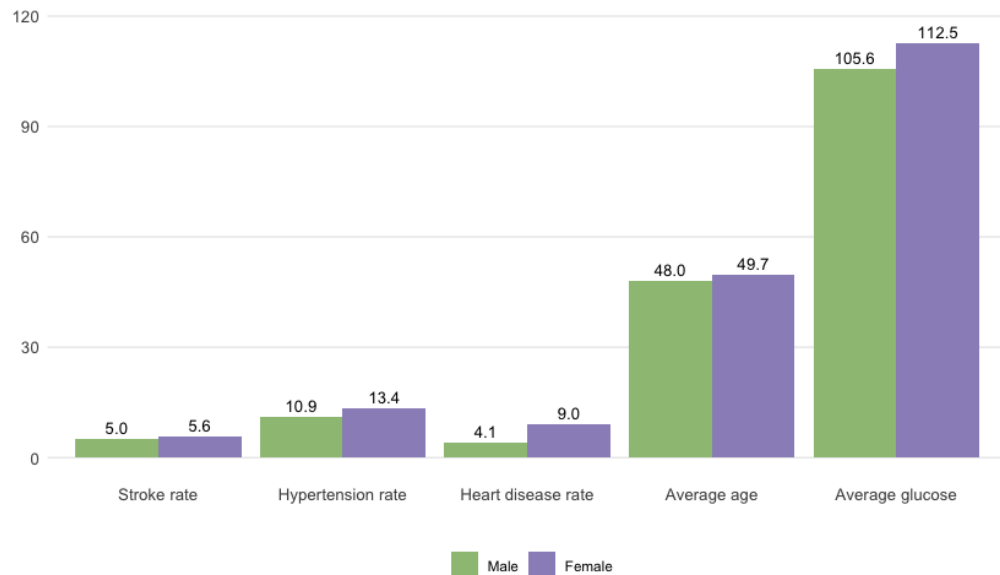
(表一) 不同性別的健康風險因子比率表

	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%

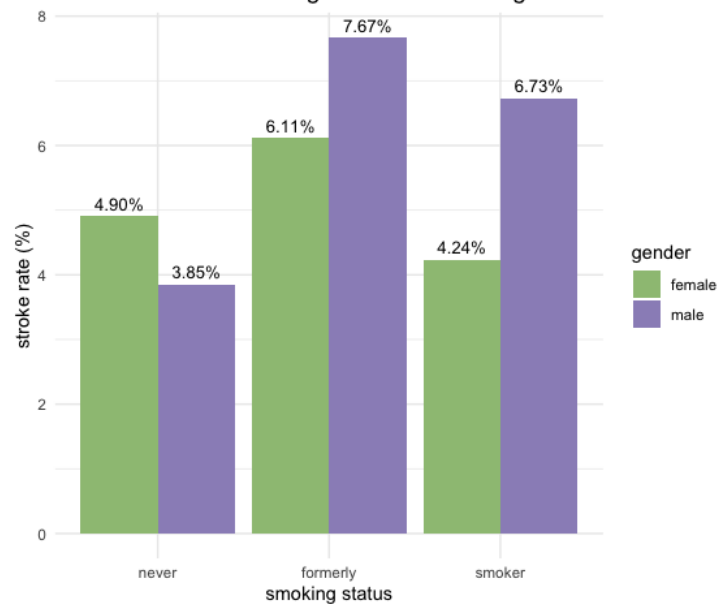
(表二) 不同性別與吸菸狀態的中風比率表

# 性別與其他風險因子的交互作用

Different gender of risk factor comparison



Stroke rate of different gender and smoking status



# 風險因子組合

基礎風險

無風險因子：3.4%

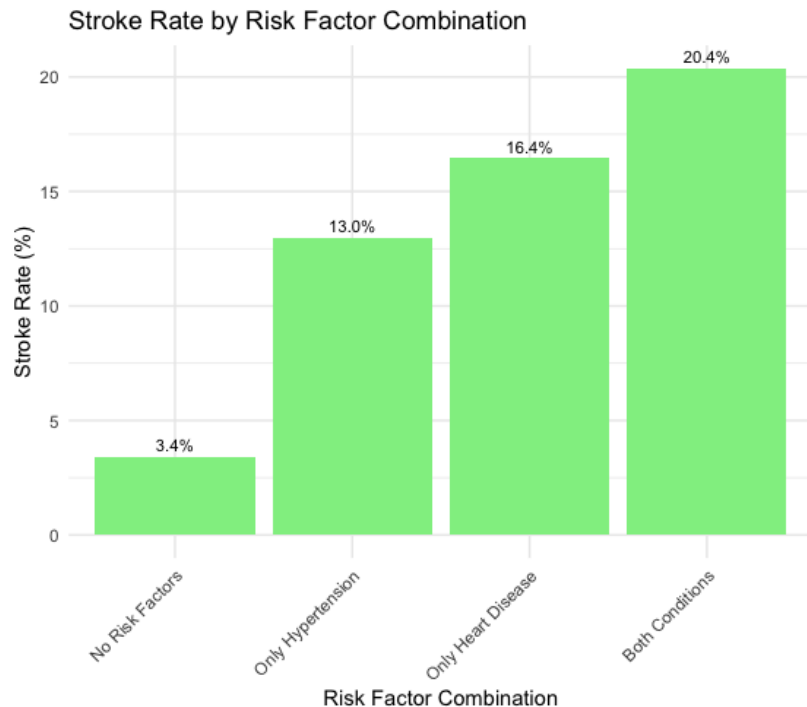
單一風險因子

僅有高血壓：13.0%

僅有心臟病：16.4%

風險因子疊加效應

同時有高血壓和心臟病：20.4%



# 年齡分層風險分析

中風發生率變化:

分層顯示隨年齡增加，中風風險持續上升。

各年齡層數據:

40歲以下風險最低，65歲以上高達13.3%。



# 風險因子影響

A red fire truck is parked on a city street at night. The truck has "DEPT. NY" and "FIRE" markings. The background shows a city street with buildings and streetlights.

## 高血壓與心臟病：

這些風險因子之間的互動值得重視，以進行有效的風險評估。

## 性別差異：

研究中發現性別也影響中風風險，在預測中不可忽視。

04-3

## 預測模型建立與評估

# 預測模型建立

## 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。

## 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。

## 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。



# 模型預測效能比較

	廣義加成模型 GAM	隨機森林模型 Random Forest	XGBoost 模型
準確率 ( Accuracy )	94.74%	94.45%	94.83%
敏感度 ( Sensitivity )	0.0182	0.0367	0.0182
特異度 ( Specificity )	99.9%	99.5%	100%
AUC值	0.783	0.736	0.771
F1 score	0.035	0.065	0.036
Balanced score	0.509	0.516	0.509
Final score	0.646	0.626	0.640

( 表三 ) 三種模型預測效能評估

# 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。表現如下：

- 準確率達94.74%，顯示模型具有優秀的整體預測準確性
- 特異度達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

# 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

# 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

# 結論

- 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。

04-4

## 模型預測準確度分析

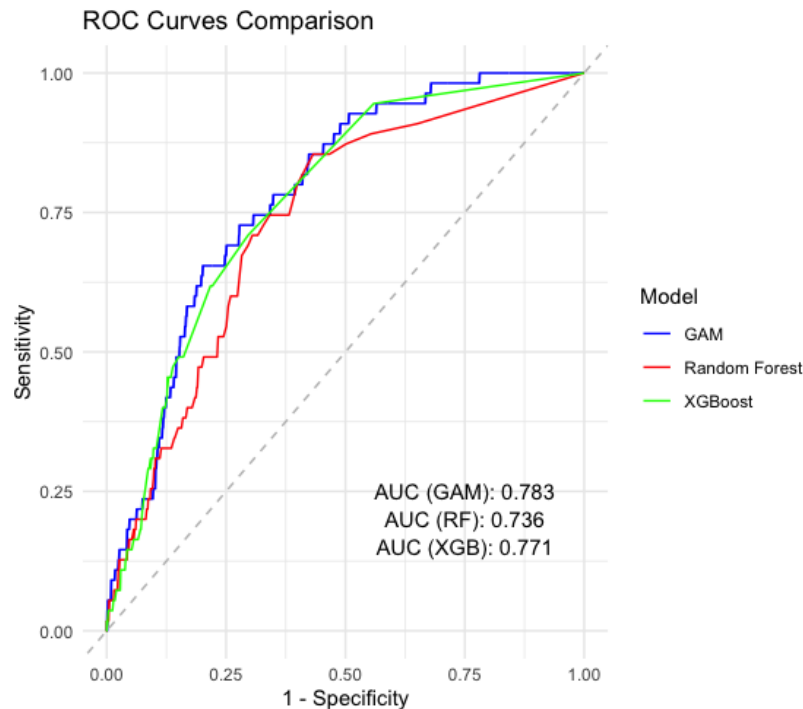
# 模型預測準確度指標

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。

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。

# ROC曲線比較模型的性能分析圖

本研究對三種模型的ROC曲線和AUC值進行了深入分析，如圖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。





# 模型效能比較

**GAM效能最佳：**

其準確率為94.74%，AUC達0.783，表現良好。

**隨機森林穩定：**

在敏感度表現優越，接近0.0367的數值。

**XGBoost特性：**

顯示出在特定情況下的優勢，AUC為0.771。

04-5

## 模型在不同族群中的預測表現

# 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。

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%）。

# 不同年齡層模型預測性能比較分析長條圖

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。



# 年齡層別預測效能

## 高齡族群 (ELDERLY, 65歲以上)

準確率：三個模型皆達到約85-90%的準確率

敏感度：較低(約5%)，顯示對實際中風案例的識別能力有限

特異度：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

## 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

## 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

05

## 結論與建議

# 研究結論

本研究使用GAM、隨機森林和XGBoost三種機器學習方法進行中風預測分析。研究結果顯示，三種模型均具有超過94%的預測準確率，其中GAM模型展現出最佳的整體預測能力，AUC值達0.783。

在風險因子分析中，研究發現年齡是最關鍵的中風風險因子。40歲以下年輕族群的中風發生率為0.8%，而65歲以上高齡族群則顯著提高至13.3%。高血壓和心臟病的協同作用也明顯增加中風風險，當這兩個風險因子同時存在時，中風風險可達20.4%。

性別分析顯示，男性在心血管疾病負擔較重，且吸菸行為對男性中風風險的影響更為顯著。研究結果建議，臨床實務可根據不同族群特性選擇合適的預測模型，以提高中風預防和風險管理的效率。

# 臨床應用建議



## 針對不同年齡層：

應用不同的模型提升預測準確度  
，針對高風險族群加強追蹤。

## 建立分級預防機制：


制定針對風險因子的預防方案，  
提升健康管理。

## 強化高危族群管理：

針對已被確定的高風險族群，進  
行後續的觀察與處理。



# 未來研究方向




## 擴大研究樣本:

將樣本數量擴大，以提高研究的外部效度。



## 納入更多風險因子:

探索其他可能影響中風的風險因子，進行更加全面的分析。



## 深度學習的應用潛力:

開發基於深度學習的新方法，進一步提高預測準確性。



**THANKS FOR  
LISTENING**