114 學年度資料探勘 第十組

MRI與阿茲海默症

第十組

組長：112029024 陳暐凱

112029044 蔡碩恩

111007050 傅景裕

日期：2026 年 1 月 12 日

目次

研究背景

本研究採用由Daniel S. 等人（2007）與 Marcus 等人（2010）發布之「開放存取影像研究系列」（Open Access Series of Imaging Studies, OASIS）資料集。該資料集提供公開的結構性磁振造影（structural MRI）之腦部影像數據，以及其他數據包括：認知能力分數、臨床失智症量分數、各項社會人口學資料（如：社經地位、教育程度、年齡）。本組將聚焦於阿茲海默症（Alzheimer's disease, AD）之早期辨識與疾病進程推估，並以MRI作為主要資料來源，輔以各項量表分數及社會人口學資料。

（一）阿茲海默症

阿茲海默症已然成為現代醫學與公共衛生領域中最嚴峻的挑戰之一。作為一種漸進式且不可逆的神經退行性疾病，阿茲海默症不僅侵蝕患者的認知功能，更對家庭與社會照護體系造成沉重的負擔。失智症患者在發病初期通常會先經歷短期的記憶缺失。然而，隨著病程進入中後期，患者的語言表達、邏輯判斷力將會逐漸喪失，最終導致其喪失獨立生活的能力。患者初期的症狀，往往被誤認為是自然老化的現象，因此在早期介入並準確辨識疾病尤顯重要。

從生物病理學的角度來看，阿茲海默症的發生與大腦內 β-類澱粉蛋白斑塊（Amyloid plaques）的沉積，以及 Tau 蛋白纏結（Tau tangles）的異

常累積密切相關，這些病理性蛋白質的堆積會觸發連鎖反應，導致神經元受損並大量死亡，在宏觀尺度上會直接反映為腦部萎縮，特別是負責記憶處理的核心區域——海馬迴（Hippocampus），以及大腦皮質區域。因此，如何精確地觀察與量化這些物理上的結構變化，便成為診斷阿茲海默症的關鍵。

（二）結構性磁振造影 MRI

而 MRI 技術正好能夠準確量測出大腦結構上的影像數據，雖然 MRI 無法直接看到神經異常蛋白的堆積，但它能捕捉疾病在大腦結構層面造成的後果，例如與記憶功能高度相關的內側顳葉、海馬迴萎縮，或是全腦體積的相對縮小。且相比於傳統的影像檢查，MRI 具備極高的組織解析度，能夠在非侵入性的情況下，提供清晰的結構細節，使研究者能精確測量大腦體積變化並追蹤腦部整體萎縮的進程。

（三）研究目的

本研究採用兩種研究資料集：橫斷研究（cross-sectional）與縱貫研究（longitudinal）資料集。

橫斷研究的研究目的是：在某一個時間點，是否能根據受試者的特徵，辨識其是否已呈現失智相關異常，此問題為分類（classification）任務，也更貼近臨床的即時篩檢情境。

而縱貫研究目的則是：在一段追蹤時間後，腦部結構會如何變化，此任務為回歸（regression）或變化量預測，重點不是當下的分類，而是將基線資訊（baseline）與追蹤間隔（follow-up delay）納入後，推估未來的腦部萎縮趨勢，為治療上提供更多可參考資料。

資料說明

（一）橫斷研究資料

橫斷面資料集涵蓋了 416 名受試者，其樣本廣度從成年早期延伸至高齡期，總計包含了 436 筆訪視數據，每一筆數據代表受試者的一次臨床訪視，而在每次訪視過程中，研究並非僅進行單次攝影，而是會連續攝取 3 至 4 次的 MRI 影像，再透過多個影像的疊加與平均，提升數據穩定及精確度。其中 20 筆為一組由 20 名健康受試者所構成的可靠度子資料集（Reliability Subset）。這些受試者在初次掃描後的 90 天內進行了第二次的掃描，其目的為評估MRI影像測量工具在短時間內的穩定性。然而，在本研究為維持資料的獨立性，並避免自相關干擾，在資料前處理階段決定剔除這 20 筆重複測量的紀錄，提升研究結果的外部效度。

受試者的年齡分佈自 18 歲的成年早期延伸至 96 歲的高齡晚期，其中共有 100 名60 歲以上的高齡族群被診斷為患有不同程度的阿茲海默症，其病程嚴重程度涵蓋極輕度到中度，其餘受試者則為正常的對照組。在性別分佈上涵蓋了男性與女性樣本。另外，本資料集所有受試者皆為慣用右手，旨在排除大腦半球側化對腦結構分析可能帶來的雜訊影響。

資料集中具臨床代表性的特徵，其具體醫學意義與統計定義如下：

臨床失智症評分（Clinical Dementia Rating, CDR）：這是本研究最重要的分類指標。CDR 是透過專業醫師對患者在記憶力、定向力、解決問題能力、社區活動、居家生活及個人照護六大維度進行評估後得出的分數。分數範圍包含健康（0）、極輕度（0.5）、輕度（1）及中度（2）。在本研究的模型設計中，依據Daniel S. 等人（2007）與 Marcus 等人（2010）文獻中的診斷標準，CDR > 0 即定義為患有失智症，是模型分類的目標特徵（Y）。

簡易心智狀態測驗（Mini-Mental State Examination, MMSE）：MMSE 是臨床上最常用的認知篩檢工具，總分為 30 分。其測試內容涵蓋時間與地點定向、語言表達、注意力與計算力等。MMSE 能有效反映患者當下的認知功能水準，測驗分數低於 24 分則具有認知功能異常，大於等於 24 分為認知功能正常。

標準化全腦體積（Normalized Whole Brain Volume, nWBV）：為大腦結構指標，透過估計顱內容積（eTIV）及影像縮放係數（ASF）校正個體差異，nWBV 的數值代表標準化過後的腦組織佔顱內總體積的百分比，能提供比絕對腦體積更具個體可比性的數據。其數值越低，代表腦組織在顱內體積佔比越低，意即腦組織萎縮越嚴重。

橫斷研究資料類型與說明如表一：

表一

變項名稱	型態	中文解釋	補充	範圍	統計摘要
ID	名目	訪視編號	每筆資料對應一次訪視（每次包含 3 – 4 張 MRI）	436 唯一值	-
M/F	名目	性別	男、女	M、F	F：約 64% M：36%
HAND	名目	慣用手	控制慣用手 降低大腦發展差異	R	全部右手
SES	次序	社經地位	綜合教育與職業等級計算1 = 最高、5 = 最低	1 – 5	眾數：3 低社經1、2略多於高社經（4、5）
AGE	連續	年齡	-	18 – 96	平均：77 中位數：78 SD=8.43 分布較對稱平均
EDUC	連續	受教育年數	年數越高 教育程度越高	6 – 23	平均：14.62 SD=2.93 分布較對稱平均
DELAY	連續	重新掃描間隔天數	評估影像測量穩定性 僅 20 名年輕（二十幾歲）受試者有資料	1 – 89 限九十天內	平均：20.55 正偏，大部分間隔短
MMSE	等距	簡易心智狀態量表	檢查記憶、注意力、語言、時間定向等認知能力，分數越高，認知越好，小於 24 視為認知能力異常	7 – 30	平均：27.41 中位數：29 明顯負偏 多正常者
CDR	等距	臨床失智症評分	用於判斷失智症的嚴重程度，0 = 正常、0.5 = 極輕度、1 = 輕度、2 = 中度、3 = 重度，大於 0 視為患有失智症	0, 0.5, 1, 2	平均數：0.31 中位數：0 正偏態 多正常者

（二）縱貫研究資料

資料集包括 150 名受試者的長期追蹤，總計累積了 373 筆訪視數據。與橫斷分析不同的是，每位受試者皆接受了兩次以上的臨床隨訪，且每次訪視之間的間隔時間嚴格設定為至少一年。因此能捕捉到受試者大腦結構與認知功能隨時間的演變軌跡。在影像獲取的品質控管上，研究維持與橫斷面研究一樣的技術流程，即每次訪視過程中皆掃描3 至 4 次 MRI 影像進行疊加與平均，以維持數據的穩定性與精確度。

受試者的年齡分佈精確鎖定在 60 歲至 96 歲之間，主要為阿茲海默症風險高的高齡群體。樣本組成在性別上維持了均衡分佈，同樣為了排除大腦半球側化對結果造成影響，確保所有受試者皆為慣用右手。

資料集共分為三大群體：72 名在追蹤期間始終保持認知健康的非失智者（Nondemented），其次是 64 名已確診罹患阿茲海默症（Demented）的患者，其病程嚴重程度橫跨極輕度、輕度至中度。另外 14 名受試者的轉變組（Converted），受試者在首次訪視時被判定為非失智，但在後續的一年或多年隨訪過程中，其臨床診斷逐漸轉變為失智狀態。

需特別說明一下，縱向研究的 Delay，與橫斷研究之意義不同，縱向研究的 Delay 代表了受試者自首次訪視（Baseline Visit）算起，至後續各次隨訪（Follow-up Visits）之間所歷經的具體天數。

縱向研究資料類型與說明如表二：

表二

變項名稱	型態	中文解釋	補充	範圍	統計摘要
SUBJECT ID	名目	受試者編號	同個受試者不同訪視次數	1-150 唯一值	-
MRI ID	名目	訪視編號	每個受試者至少兩次訪視，每次訪視會有一個不同MRI ID	1 – 373 唯一值	-
VISIT	名目	訪視次序	訪視第幾次 每位至少 2 次訪視	1 – 5	眾數為二
M/F	名	性別	男、女	M/F	F：60%、M：40%
HAND	名	慣用手	控制左右腦差異	R	全部右手
GROUP	名目	組別	NONDEMENTED：一直正常， DEMENTED：首次訪視便確診， CONVERTED：正常後變確診	NONDEMENTED DEMENTED CONVERTED	NONDEMENTED：74 DEMENTED：64 CONVERTED：14
SES	次序	社經地位	1 = 最高；5 = 最低	1 – 5	眾數：3 低略多於高
MMSE	等距	簡單心智狀態量表	≥24 視為正常	0 – 30	平均：約26 中位數：約28 明顯負偏（健康
CDR	等距	臨床失智症評分	0 = 正常、0.5 = 極輕度 1 = 輕度、2 = 中度	0、0.5、1、2	平均：約 0.45 中位數：0 正偏（健康者多）
DELAY	連續	訪視間隔天數	訪視相隔至少一年	300 – 1500	平均：約 450 – 550 正偏 大部分間隔較短

變項名稱	型態	中文解釋	補充	範圍	統計摘要
ETIV	連續	估計顱內容積	頭顱空間大小 數字越大，頭越大	約 1100 – 2000	平均：約 1470 SD≈140 正偏 頭顱容量偏小者略
ASF	連續	影像縮放係數	係數越大，頭越大	約 0.85 – 1.6	平均：約 1.20 SD≈0.12 正偏、頭較小者略多
NWBV	連續	標準化全腦體積	數值越低，腦萎縮越嚴重	約 0.60 – 0.90	平均：約 0.77 SD≈0.065 負偏 多數體積較大、健康

視覺化分析

在正式進入資料前處理與模型建構之前，先透過視覺化分析，初步檢視橫斷研究資料中各變項彼此之間的關聯性，觀察各變項是否與失智狀態具有潛在關聯，熱力圖結果如圖一：

圖一

分析結果顯示，臨床失智評估量表（CDR）與多項指標存在顯著關聯。最為突出的觀察在於 CDR 與簡易智能測驗（MMSE）之間呈現強烈

的負相關（$r \approx -0.75$），此數據與臨床診斷高度同步，失智症與認知功能受損程度有高度相關。

在生理結構指標上，標準化全腦體積（nWBV）與 CDR 同樣呈現中度負相關（$r \approx -0.61$），代表大腦萎縮與失智程度的正向連結：即當腦組織佔顱內空間比例下降時，受試者被診斷為失智的風險隨之升高。此外，年齡（Age）與 nWBV 之間顯著的負相關（$r \approx -0.87$），則顯示人類大腦隨老化自然萎縮的生理軌跡。

估計總顱內體積（eTIV）與影像縮放係數（ASF），與 CDR 的相關性極低，因ASF 與 eTIV 均屬於影像標準化過程中的校正參數，而非病理特徵。

Educ（教育）、SES（社經地位）與CDR、MMSE、nWBV皆呈現中低度相關（$|r| < 0.25$），代表社會人口學因素對失智症、認知功能、腦部體積較直接線性關係較低，而M/F（性別）與同樣也與上述各項變項相關性低（$|r| < 0.1$），性別並非造成個體上的差異主因。

資料前處理 橫斷研究

在進行模型建立之前，將仙針對橫斷研究資料進行資料前處理，以確保後續分析結果具有統計上的合理性與模型解釋上的清楚性。

（一）欄位刪除

首先針對不具分析價值或可能造成解釋混淆的變項進行刪除。第一個是Hand（慣用手）欄位，因其在所有受試者中皆為右手（R），不存在任何變異，因此無法提供任何區辨資訊，故直接移除。

第二個刪除欄位為 eTIV 與 ASF 兩項變數，先前的相關性熱力圖分析顯示，eTIV 與 ASF 與主要臨床指標（如 CDR、MMSE、nWBV）之線性相關程度極低，雖然與 MRI 影像處理有關，但其本質屬於影像標準化與頭顱大小校正參數，而非反映疾病狀態或腦部病理變化的指標。表示其對失智症判別的直接貢獻甚微。為避免模型引入偏誤，選擇將此兩項變數排除。

最後一個欄位為 Delay（MRI 重新掃描間隔天數），此變項目的在於評估 MRI 影像量測的穩定性，而非描述個體的腦部結構或疾病狀態。因此，Delay 與失智症本身不存在生物學上的直接關聯，亦可能干擾模型對真正病理變化的學習，故同樣予以刪除。

（二）可靠度子資料集處理

原始資料集額外包含一組由 20 名年輕健康受試者所構成的「可靠度子資料集」。此子資料集的設計目的，是讓同一名受試者在短時間（90 天內）接受兩次 MRI 掃描，以評估影像量測的一致性。此資料極存在明顯的非獨立性問題：同一個體的重複測量違反了樣本獨立原則。若將此類資料納入分析，可能導致模型在訓練與測試過程中，間接看到同一受試者的資訊，進而影響模型結果，因此選擇刪除這 20 筆資料，留下416 筆彼此獨立的樣本。

（三）遺漏值處理

接著下個步驟著手檢視各變項的遺漏值情形。結果顯示，性別（M/F）、年齡（Age）、CDR 與 nWBV 均無遺漏值；然而，教育年數（Educ）、社經地位（SES）與 MMSE 則存在大量缺失，遺漏筆數超過樣本數的一半。若採取完全刪除含遺漏值樣本的方式，將導致樣本數大幅下降，並嚴重影響資料的年齡結構，因這些缺失多數來自年輕（四十歲以下）且認知功能正常的受試者。不僅會降低模型穩定性，也可能使研究樣本偏向高齡族群，進而影響模型的泛化能力。

基於上述考量，本研究採取雙策略設計，將資料分為兩個版本進行比較：

資料集 A（補值策略）：保留 Educ、SES 與 MMSE，並對其遺漏值進行補值處理，根據前述熱力圖相關性分析結果，Educ 與 SES 與主要臨床指標的線性相關程度偏低，對於模型影響小，因此直接採用中位數補值。至於 MMSE，原始資料說明文獻指出，其遺漏情形出現在認知功能正常者（MMSE >= 24），但並未提供每筆缺失資料的實際分數。因此同樣選擇以整體樣本的中位數（29 分）進行補值，使補值結果維持在合理且正常認知功能的範圍內。

資料集 B（刪除變項策略）：直接移除 Educ、SES 與 MMSE 三項變數，僅保留完整且無遺漏的基本特徵。

（四）異常值檢查

對於連續型變項，本研究採用 IQR 法則 進行異常值檢查，設定上下界為 Q1 − 1.5×IQR 與 Q3 + 1.5×IQR。結果顯示，所有連續變項均未出現超出合理範圍的極端值。

類別型與次序型變項（如性別、CDR、SES）亦皆符合其原始定義範圍，未出現不合法值，無需進一步進行異常值剔除。

（五）分類編碼 One - Hot Encoding

首先類別變項需轉換為模型可接受的數值形式：將性別由文字型態轉換為二元數值編碼（男性為 1、女性為 0）。

接著根據 Daniel 等人（2007）之原始文獻，本研究將 CDR 分數轉換為二元分類目標變數 Dementia：CDR = 0 視為正常，將其編碼為 Dementia = 0；CDR > 0 視為失智症患者，將其編碼為 Dementia = 1。Dementia即為橫斷研究是否罹患失智症的分類目標。

（六）特徵選擇

本組第一個研究目的是透過橫斷資料，建立一個能夠根據個體的結構性 MRI 數據與臨床背景資訊、輔以社會人口學資料，判別是否罹患阿茲海默症的分類模型，以此研究目的為前提，將模型的輸出特徵 Y 設定為 Dementia：Dementia = 0 代表正常、無失智症，資料數為 316 筆，Dementia = 1 異常、有失智症，資料數為100筆，（0 / 1）比例為 3.16。

輸入特徵 X 的部分，Model A（補值）的特徵選擇有 M/F、Age、nWBV、Educ、SES、MMSE，而 Model B（直接刪除）特徵有 M/F、Age、nWBV。 年齡與性別作為基本人口學背景變項，提供模型對自然老化與性別差異的基準調整，nWBV 則作為腦部萎縮的核心影像指標，MMSE 與 CDR 則分別代表量化認知表現與失智症診斷嚴重度。

實驗設計 橫斷研究

此實驗採用分層五折交叉驗證（Stratified 5-fold Cross-Validation）結合 SMOTE 過採樣的設計，以同時處理資料不平衡問題並確保模型評估的穩定性。

首先，整體資料集依照目標變數 Dementia（是否失智）進行分層抽樣，將資料劃分為五個大小近似的子集合（folds）。在每一個 fold 中，失智與非失智個案的比例皆與母體資料一致，原始比例約為 3.16 : 1。在交叉驗證流程中，每一輪會輪流選取其中一個 fold 作為測試集（Test Set），其餘四個 folds 作為訓練集（Training Set），讓每一筆資料皆有機會作為測試樣本，充分利用資料資訊，避免浪費珍貴的臨床觀測值。

測試集維持原始真實分布，不進行任何過採樣或資料調整，保證模型是面對實際臨床資料的分布情境。接著針對每一輪的訓練集執行 SMOTE 過採樣。透過合成少數類別樣本，使訓練集中無失智症 / 有失智症（0 / 1）的比例由原始的 3.16 : 1 調整為 1 : 1（如表三）以解決失智個案數量明顯較少所造成的類別不平衡問題，降低模型偏向多數類別的風險，提升模型的學習能力。

表三

Fold	原始訓練集 (0 / 1)	SMOTE 後訓練集 (0 / 1)	測試集 (0 / 1)
1	252 / 80	252 / 252	64 / 20
2	253 / 80	253 / 253	63 / 20
3	253 / 80	253 / 253	63 / 20
4	253 / 80	253 / 253	63 / 20
5	253 / 80	253 / 253	63 / 20

建立模型 橫斷研究

（一）模型選擇

橫斷研究的分類模型選擇為隨機森林（Random Forest），因其能捕捉非線性關係，適合處理的大腦結構指標、年齡（Age）對失智症風險的影響，由於人類神經構造，這些變項間的關係，通常不是單純線性關係。

第二個原因是具有良好的穩健性。由於模型是由多棵決策樹所構成，最終預測結果是採是投票機制產生，可以避免部分樹因雜訊或局部樣本偏差，而導致整體結果錯誤。在這個抗噪能力前提之下，隨機森林在訓練階段搭配 SMOTE 方法，以平衡訓練資料類別不平衡問題（無失智症樣本數顯著多於有失智症樣本），

最後由於隨機森林能夠提供特徵重要性，透過量化各輸入變項對模型預測結果的相對貢獻程度，得知影響模型結果的重要特徵為何，本研究目的不僅在於建立一個具有良好分類效能的模型，更希望能理解年齡、腦體積指標與臨床量表（如 MMSE）在失智症辨識中的角色。從而提升模型在醫學研究中的解釋價值。

（二）模型超參數設定

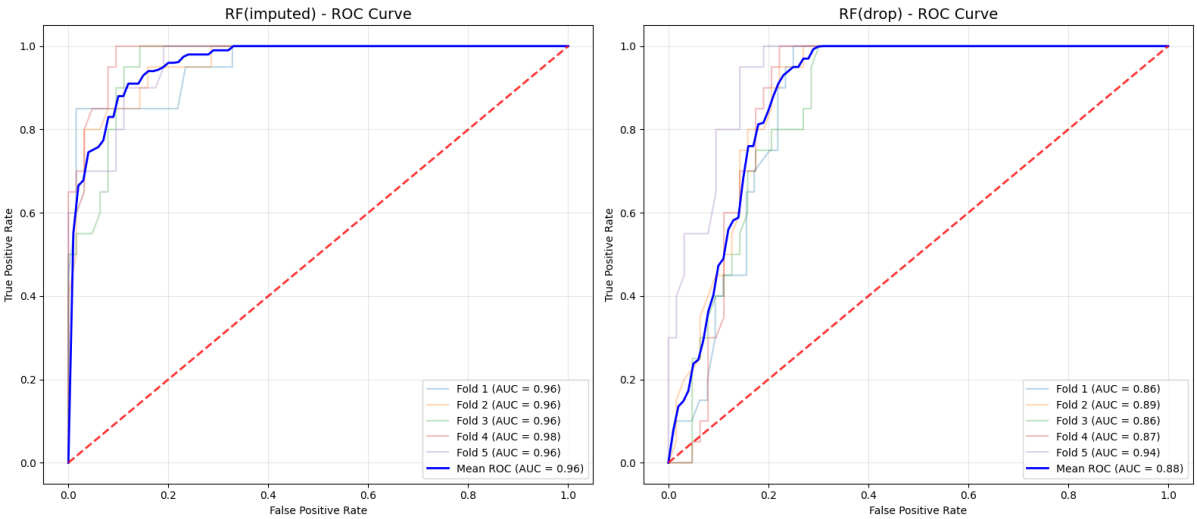
在模型超參數的設定，將樹的數量（n_estimators）設定為 100，以在模型穩定性與計算效率之間取得平衡、亂數種子（random_state）設為

42，每棵樹的最大深度不加限制（max_depth=None），使模型能充分學習資料中的潛在結構、節點分裂所需的最小樣本數（min_samples_split）設為 2，而葉節點所需的最小樣本數（min_samples_leaf）設為 1，以維持模型的表達能力。分裂準則採用 Gini impurity（criterion='gini'），此為分類問題中常用且的指標。最後，啟用 bootstrap 抽樣（bootstrap=True），使每棵樹在訓練時使用隨機抽樣且可重複的樣本子集，進一步降低模型的變異性並提升泛化能力。

模型結果解釋 橫斷研究

（一）ROC 曲線分析

圖二（Model A 補值）　　圖三（Model B 刪除）



從 ROC 曲線、 AUC 指標來看，Model A （如圖二）的平均 AUC 約為 0.906 ± 0.023，整體數值較高，各折之間的 AUC 也相對集中，顯示模型在不同資料切分下表現皆較穩定且表現良好，代表 Model A 對於失智症與非失智症個體具有良好的區辨能力，不易波動。

相較之下，Model B （如圖三）的平均 AUC 約為 0.779 ± 0.104，不僅整體辨識能力明顯下降，各折之間的 AUC 落差也顯著增大，模型表現變不穩定，對失智症風險的辨識能力明顯受限。透過比較兩模型，可以得知在保留 MMSE、SES 等臨床與社會人口學變項的情況下，模型能同時利用腦結構指標與臨床量表資訊，提升整體判斷的可靠度。

（二）分類指標分析

表四（Model A）

Fold	Accuracy	Recall	Precision	F1-score
1	0.881	0.850	0.708	0.773
2	0.892	0.850	0.739	0.791
3	0.880	0.800	0.727	0.762
4	0.928	0.850	0.850	0.850
5	0.880	0.800	0.727	0.762
平均	0.892	0.830	0.750	0.788

表五（Model B）

Fold	Accuracy	Recall	Precision	F1-score
1	0.774	0.750	0.517	0.612
2	0.819	0.750	0.600	0.667
3	0.807	0.700	0.583	0.636
4	0.807	0.650	0.591	0.619
5	0.880	0.800	0.727	0.762
平均	0.817	0.730	0.604	0.659

進一步從分類指標角度分析，Model A （如表四）的 Recall 約為 0.830，能成功辨識出大多數具有失智症的個體，顯示其對於真陽性具敏感性。雖然其 Precision 約為 0.750，代表仍存在部分假陽性（將健康個體誤判為有病），但在醫療應用情境中，特別是早期篩檢階段，這樣的取捨是

可以接受的。相較於漏診潛在病患，適度提高敏感度、多抓出可疑個案，反而更符合臨床實務對早期篩檢工具的期待。

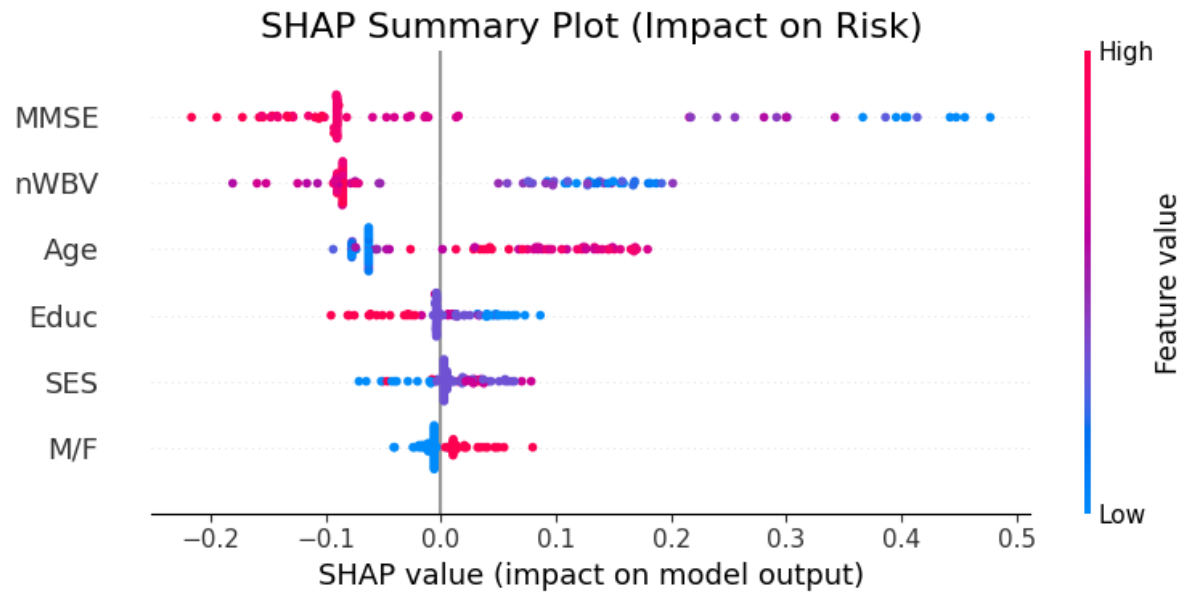
反觀 Model B（如表五），其 Recall 僅約為 0.730，模型 B 在排除 MMSE 與 SES 等資訊後，對於部分早期或症狀較輕微的失智個體辨識能力明顯下降。這意味著，Model B 在實際應用中可能會錯過一部分真正需要進一步醫療評估的個體，降低其作為早期篩檢工具的實用性。

綜合 ROC 分析與分類指標結果可以得出結論：Model A 不僅整體效能較佳，且在穩定性與敏感度方面表現更為理想，較適合作為早期失智症篩檢模型；而 Model B 則因資訊不足而導致辨識能力與穩定性下降，較不適合用於此研究目的。

（三）Model A SHAP 值分析

接著透過分析 SHAP 值，去探討各個特徵對於模型判斷失智症的影響結果如圖四所示：

圖四



模型的決策主要由 MMSE、nWBV 與 Age 三個變項所主導。最具影響力的是 MMSE，因其 SHAP 點雲呈現最寬的橫向分布，是 Model A 中最具關鍵性的決策特徵。從顏色與方向性來看，低 MMSE 分數（藍色點）明顯集中在 SHAP 值為正的區域，代表當 MMSE 分數越低時，模型會強烈將個體推向失智症的預測結果（反向）。這意味著，即使在多變項同時存在的情況下，MMSE 仍然是模型判斷失智風險時最具權重的臨床指標。

標準化全腦體積（nWBV）是影響力排名第二的， 其 SHAP 分布寬度僅次於 MMSE，腦體積較小（藍色點）多分布於 SHAP 值為正的區域，表示當 nWBV 越低、腦萎縮程度越高時，模型越傾向預測為失智症，也是呈現反向，這個結果與神經退化性疾病的病理機制一致。

影響力第三重要的是年齡（Age），但其影響程度低於 MMSE 與 nWBV。SHAP 圖顯示，高齡個體（紅色點）主要分布在 SHAP 值為正的區域，表示隨著年齡增加，模型對失智風險的預測機率也隨之上升，屬於正向影響，這與失智症多為年齡長者的臨床情況相符。

接著是教育程度（Educ）與社經地位（SES），SHAP 點雲分布相對狹窄，且多集中在 SHAP 值接近 0 的區域，代表兩變項對模型最終判斷的影響有限。性別（M/F）的影響力幾乎沒影，SHAP 值幾乎完全貼近 0，點雲近乎收斂成一條直線，代表其對失智症幾乎不具預測力

（三）綜合分析

將 SHAP 分析結果與前述熱力圖及 Model A、B 的比較結果整合後，MMSE、nWBV 與 Age 為模型中最主要的三個決策特徵，不僅在 SHAP 分析中具有最高影響力，在相關性分析中亦與 CDR 呈現顯著關聯。

總體結果來說，MMSE 分數越低，CDR 異常的機率越高、代表腦體積萎縮程度越高，失智風險越大、而 Age 與 CDR 呈正相關，年齡上升伴隨失智風險增加。

反之，Educ、SES 與性別在熱力圖中與 CDR 的相關性本就較低，SHAP 分析亦顯示其對模型輸出的影響有限，進一步說明這些變項並未對分類結果提供關鍵資訊。這也解釋了為何在 Model B 中刪除 MMSE、

Educ 與 SES 後，模型效能與穩定性會明顯下降，特別是對早期與輕度失智個體的辨識能力受到影響。

從應用角度來看，Model A 的預測行為呈現出高敏感度的特徵，符合臨床早期篩檢工具的設計目標。相較於追求極高精確率的診斷模型，早篩模型更重視能否及早辨識潛在患者，再交由後續更精細的臨床檢查確認。因此，該模型具備應用於高齡族群健康檢查、或作為初步風險分流工具的潛力。

未來優化方向，可由資料視覺化分析提供的線索著手。結果顯示，Age 與 nWBV 之間存在高度負相關（$r \approx -0.87$），反映年齡增加與腦體積萎縮之間的強烈關聯，兩者不僅各自影響失智風險，其兩變項交互作用本身也可能攜帶重要資訊。未來研究可考慮在模型中顯式加入 Age × nWBV 的交互特徵，以捕捉不同年齡層中腦萎縮對失智風險影響程度的差異，進一步提升模型對早期與邊界個案的辨識能力。

資料前處理 縱貫研究

（一）欄位刪除

由於橫斷研究結果顯示，部分人口學與結構性變項對模型辨識能力貢獻有限，因此在縱貫分析中，刪除了 MRI ID、性別（M/F）、慣用手（Hand）、教育程度（Educ）、社經地位（SES）、顱內容積（eTIV）與 ASF 等欄位，以去除冗餘資訊與模型複雜度，僅保留Subject ID（僅作為個體識別）、Visit（用於界定追蹤次序）、MR Delay（每次 MRI 測量間隔天數）、Age、MMSE、CDR 與 nWBV。

（二）遺漏值、異常值處理

經檢查後發現僅有 MMSE 存在兩筆缺漏資料。由於這兩筆資料在時間序列中缺乏前後對照資訊，無法合理推估其數值，為避免引入人為偏差，本研究選擇直接將該兩筆樣本剔除，剩下 371 筆有效資料。接著透過描述性統計與異常值檢查，確認各數值變項皆落於合理範圍內，未發現需額外處理的極端值或異常分布。

（三）資料重組配對

為符合縱貫研究一人一筆的建模設定，將資料轉換為配對形式，取每位受試者 Visit = 1 的資料作為 baseline（基準測量），並選取該受試者 Visit 最大值所對應的資料作為 follow-up（追蹤測量）。接著，將 baseline

與 follow-up 的資訊合併為單一觀測單位（如表六），使每位受試者僅保留一筆資料，以避免重複測量造成的樣本相依性問題，最終可用於回歸分析的樣本數為 150 筆。

表六

Subject ID	Visit_followup	MR_Delay_followup	Age_followup	MMSE_followup	CDR_followup	nWBV_followup
OAS2_0001	2	457	88	30.0	0.0	0.681
OAS2_0002	3	1895	80	22.0	0.5	0.701

Subject ID	Visit_baseline	MR_Delay_baseline	Age_baseline	MMSE_baseline	CDR_baseline	nWBV_baseline
OAS2_0001	1	0	87	27.0	0.0	0.696
OAS2_0002	1	0	75	23.0	0.5	0.736

（四）特徵選擇

根據研究目的二：預測未來病症發展，將問題定義為回歸任務，目標為推估受試者在未來某一追蹤時間點的全腦體積比例（nWBV_followup）。因此，模型輸入特徵 X 僅採用在 baseline 時即可取得、且具臨床解釋意義的變項，包括 Age_baseline、MMSE_baseline、CDR_baseline、nWBV_baseline，並額外納入 MR Delay_followup 以反映不

同受試者之追蹤時間長短，而模型的輸出目標Y則設定為

nWBV_followup，表七為特徵資料分布情況：

表七

變項	Mean	SD	Min	Median	Max
Age_baseline	75.46	7.57	60	75	96
MMSE_baseline	27.57	2.98	17	29	30
CDR_baseline	0.26	0.33	0	0	1
nWBV_baseline	0.736	0.037	0.660	0.736	0.837
MR_Delay_followup (days)	1068.29	541.55	365	846	2639
nWBV_followup	0.723	0.038	0.644	0.721	0.827

實驗設計 縱貫研究

縱貫實驗設計延續橫斷研究的策略，採用 5-fold 交叉驗證（cross-validation） 作為資料拆分與模型評估的方法，將最終整理完成的一人一筆縱貫資料集，隨機劃分為五個大小近似的子集合，每一折（fold）皆輪流作為測試集，其餘四折則作為訓練集（如表八），重複進行五次，使每一筆資料皆曾一次被用於模型測試。

表八

Fold	訓練集	測試集
Fold 1	120	30
Fold 2	120	30
Fold 3	120	30
Fold 4	120	30
Fold 5	120	30

建立模型 縱貫研究

（一）梯度提升回歸（Gradient Boosting Regressor, GBR）選擇理由

縱貫研究選擇使用梯度提升回歸（Gradient Boosting Regressor, GBR） 作為主要預測模型，以預測受試者於追蹤時間點的全腦體積比例（nWBV）。原因與橫斷研究雷同，因腦體積隨時間變化的過程，並非單純線性關係，年齡、認知功能（MMSE、CDR）與腦萎縮之間，也可能存在非線性效應與變項之間的交互作用。GBR 透過逐步疊加多棵決策樹的方式，能夠有效捕捉這類複雜的非線性關係，而不需事先假設資料符合線性或常態分布。另外因縱貫資料最終為 150 人，梯度提升回歸相較於深度學習等高參數模型，能透過逐步修正前一輪模型殘差的方式，對於小樣本具有較佳的穩定性與泛化能力，同時降低對雜訊過度擬合的風險。再加上模型亦可計算特徵重要性，有助於後續分析，提升模型的可解釋性。最後較實際的考量為 GBR 需要的前處理要求相對寬鬆，負擔較低。

（二）模型超參數設定

本研究設定隨機種子 random_state = 42。樹的數量（n_estimators）設定為 300，使模型能透過多次迭代逐步學習殘差結構；學習率（learning_rate）設定為 0.05，降低單棵樹對整體模型的影響，促進更穩定的學習過程；最大樹深（max_depth）限制為 3，以避免決策樹過於複

雜而記憶訓練資料，降低過擬合風險；此外，透過 subsample = 0.9 的設定，每次僅隨機抽取 90% 的訓練資料進行學習，引入隨機性以進一步提升模型的泛化能力。

模型結果解釋 縱貫研究

（一）指標分數分析

縱貫研究模型指標分數結果如表八所示：

表九

Fold	RMSE	MAE	R Square	Pearson r
Fold 1	0.016647	0.012598	0.846774	0.926247
Fold 2	0.015784	0.011782	0.813863	0.909814
Fold 3	0.011698	0.008682	0.907170	0.955325
Fold 4	0.013393	0.010691	0.837455	0.929734
Fold 5	0.013799	0.011196	0.874457	0.938699
---	---	---	---	---
平均值	0.014264	0.010990	0.855944	0.931964
標準差	0.001972	0.001472	0.035929	0.016729

平均決定係數（R Square）達到 0.856，代表模型能夠解釋約 85.6% 的未來全腦體積比例（nWBV_followup）變異量，代表模型已成功捕捉到影響腦體積變化的關鍵資訊，對於未來腦萎縮趨勢具備高度的預測能力。

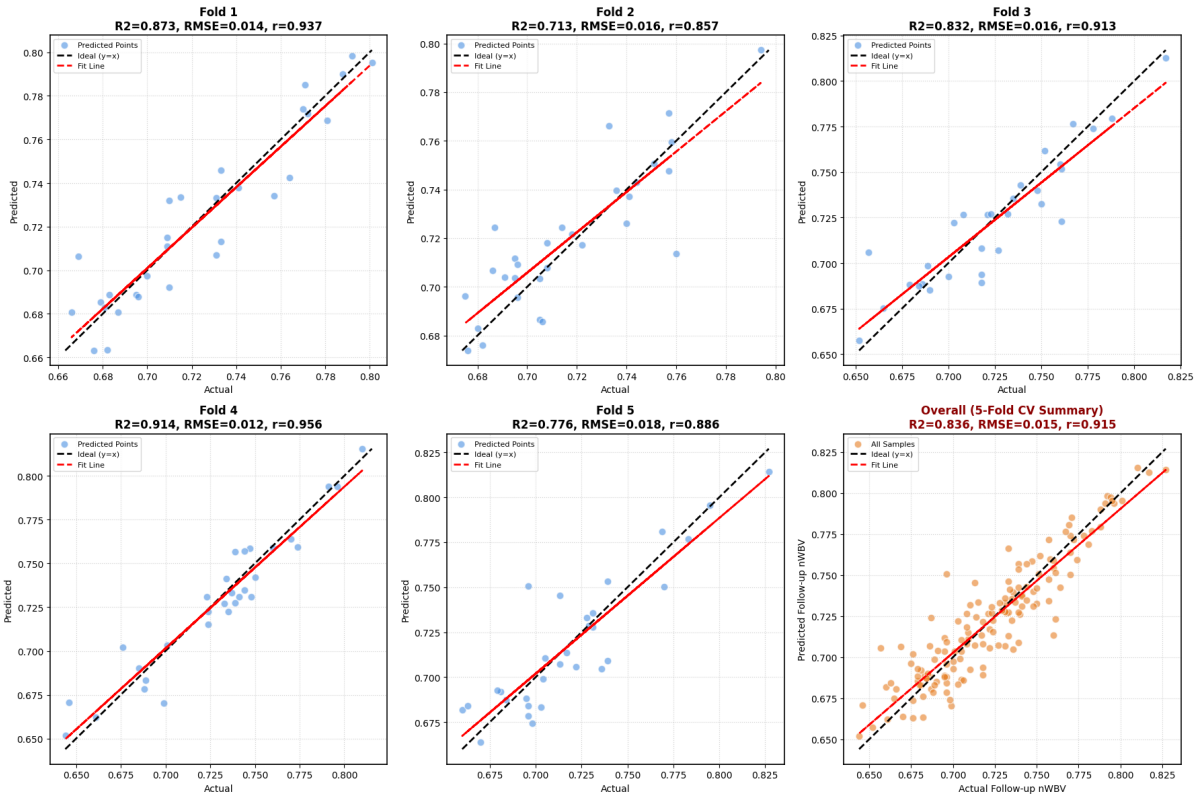
在預測誤差方面，模型的平均 RMSE 約為 0.014，MAE 約為 0.011，這兩項指標分別反映預測值與實際觀測值之間的平方誤差與絕對誤差，其數值均落在 0.01 左右，顯示模型在實際數值層面上的預測偏差相當小。平均 Pearson 相關係數達 0.932，呈現極強的正相關關係。此結果代表模型所預測的腦體積變化方向，與實際發展趨勢高度一致。

最後，無論是 R Square、RMSE、MAE 或 Pearson 相關係數，各項指標在 5-fold 交叉驗證下的標準差皆維持在極低水準，代表模型並未過度依賴特定子樣本，也未出現明顯的過度擬合現象，能維持穩定預測能力，展現出良好的泛化性。

（二）回歸散布圖

將實際腦體積設定為 X 軸，預測腦體積設定為 Y 軸，畫出以下回歸散布圖（圖五）：

圖五



圖中每一個子圖分別對應一個 fold，右下角則為整體彙總結果，黑色虛線為理想情況下的對角線（y = x），代表預測值與實際值完全一致；紅色實線則為實際資料點所擬合出的回歸線。從結果可以觀察到，散布點大多緊密分布在對角線附近，顯示模型在不同 fold 中皆能準確預測未來腦體積的數值大小，且預測誤差整體偏小，模型並未出現系統性高估或低估的情形，即便在個別 fold 中，彼此R^2 值差異也恨小，整體仍維持良好穩定解釋力。

（三）特徵重要度

下表（表十）為特徵重要性結果：

特徵名稱	特徵重要度	百分比 (%)
nWBV_baseline	0.919085	91.91%
followup_days	0.041346	4.13%
MMSE_baseline	0.022621	2.26%
Age_baseline	0.013397	1.34%
CDR_baseline	0.003551	0.36%

nWBV_baseline 為模型中最主要的預測因子，其重要度約為 0.92（91.9%），遠高於其他變項，因為未來的腦體積變化，正是將個體當下的腦體積狀態作為基準而變化。換言之，若已知基準時點的全腦體積，在預測後續腦萎縮程度時，能獲得最具解釋力的結果。

第二名是追蹤天數（followup_days）為第二重要的預測因子（約4.13%）。反映時間在腦萎縮累積過程中的關鍵角色。隨著追蹤間隔拉長，腦體積變化的幅度通常也隨之增加，

其他特徵如MMSE、Age 與 CDR_baseline 的重要度皆低於 3%，屬於輔助性預測變項。這些變項雖然在臨床上與失智風險及疾病嚴重度高度相關，但在同時納入基準腦體積的情況下，其額外提供的資訊多為間接或次級影響。

參考文獻

Daniel S., Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. 2007. “Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults.” Journal of Cognitive Neuroscience19(9)(September):1498 – 1507.doi:10.1162/jocn.2007.19.9.1498.

Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci. 2010 Dec;22(12):2677-84. doi: 10.1162/jocn.2009.21407. PMID: 19929323; PMCID: PMC2895005.

G. A. Ansari, Sivakani. R, S. Srisakth.(2022).Precise diagnosis of alzheimer's disease using recursive feature elimination method.Int. J. Systematic Innovation.7(3):28-38.

https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers/data

https://www.researchgate.net/figure/A-C-Analysis-of-ASF-eTIV-and-nWBV-for-Demented-and-Non-demented-group_fig5_359412479

https://matilda.fss.uu.nl/articles/common-data-types.htmlMarcus,

附錄

（一）組內分工表

工作內容	負責人
資料、文獻查找	全員
研究背景	傅景裕
資料說明	傅景裕
視覺化分析	傅景裕、陳暐凱
資料前處理 橫斷研究	全員
實驗設計 橫斷研究	傅景裕、陳暐凱
建立模型 橫斷研究	傅景裕、陳暐凱
結果解釋 橫斷研究	傅景裕
資料前處理 縱貫研究	全員
資料前處理 縱貫研究	傅景裕
實驗設計 縱貫研究	傅景裕
建立模型 縱貫研究	傅景裕
結果解釋 縱貫研究	傅景裕
期中簡報製作	全員
期末簡報製作	傅景裕
書面報告	傅景裕

（二）Q & A

提問一：橫斷面與縱向研究的區別是什麼？
111007050 傅景裕：橫斷面年齡橫跨18至96歲，跨度大，此數據顯示僅六十歲以上高齡者，較容易患有失智症，故縱向研究僅針對六十歲以上較

提問二：橫斷面與縱向研究的ＩＤ各自代表什麼？沒有意義是否可以刪除？
111007050 傅景裕：這些ＩＤ僅代表研究進行中的受試者訪視編號，沒有任何實質統計的必要，在後續資料前處理的部分會刪除
提問三：MMSE與CRD分數與失智症間有什麼關係？
112029024 陳暐凱：前者代表認知能力分數，後者代表失智症嚴重分數，兩者呈現高度正相關，因此在前處理的過程，會將以CRD作為判斷失智症與否的指標
提問四：退行性疾病是指什麼？
111007050 傅景裕：退行性特指疾病只會越來越糟糕，且疾病狀況是不可逆的，失智症便屬於此類疾患
提問五：這些年齡與 nWBV 負相關、CDR 與 MMSE，是符合正常現象嗎？
111007050 傅景裕：由於人類腦組織本就會隨著時間，漸漸凋亡萎縮，屬自然老化現象，隨著年齡增長，腦體積便會下降，意即 nWBV 下降，與年齡呈負相關合理；而MMSE 用以測量認知能力，CDR 量表測量項目同樣包含認知能力，故兩量表之間會有相關，但 MMSE 量表分數越高，認知能力越好， CDR 量表則是分數越低越正成，故兩者呈現負相關，也屬合理
提問六：eTIV、ASF 兩個變項為什麼要刪除不計？
111007050 傅景裕：eTIV 與 ASF 與主要臨床指標（如 CDR、MMSE、nWBV）之線性相關程度極低，雖然與 MRI 影像處理有關，但其本質屬於影像標準化與頭顱大小校正參數，而非反映疾病狀態或腦部病理變化的指標。表示其對失智症判別的直接貢獻甚微。為避免模型引入偏誤，選擇將此兩項變數排除。
提問七：要怎麼橫斷資料正常、異常比例失衡的問題？

111007050 傅景裕：針對每一輪的訓練集執行 SMOTE 過採樣。透過合成少數類別樣本，使訓練集中無失智症 / 有失智症（0 / 1）的比例由原始的 3.16 : 1 調整為 1 : 1（如表三）以解決失智個案數量明顯較少所造成的類別不平衡問題，降低模型偏向多數類別的風險，提升模型的學習能力。
提問八：縱貫研究回歸模型的部分，R Square與皮爾森相關係數各是多少？
111007050 傅景裕：R Square 約為 0.856、皮爾森相關係數約為 0.932

（三）討論紀錄

資料探勘第十組 討論紀錄 LINE文字檔 谷歌雲端連結

https://drive.google.com/drive/folders/1GGr1JrEDtLB5_xwCeIK65WXU5nG3fkjk?usp=sharing

（四）個人心得

111007050 傅景裕：

對於一個心理系的外系生來說，這門課的難度實在是頗高的，加上這次參與了專案從頭到尾的製作，老實說負擔起蠻沈重的，畢竟是個蠻陌生的領域，要將一個毫不熟悉的資料集，昇華轉換成有意義、可操作的分類、預測模型，要顧慮的層面非常非常多。在一開始，為了理解這兩個資料集的收集背景與動機，特別去查閱了原文獻的說明，在腦中漸漸形成初步的研究方向。接著進行資料集資料說明的時候，試著練習搞懂每個變項

所代表的含義、資料分布情況，以利後續的資料清理、篩選等等，也是挑戰的開始..

來到了資料前處理的部分，要面對的是一大堆的遺漏值，要學習如何處理缺失值，才能讓資料既保持原來的分布情況，不影響訓練模型的結果，同時也要有足夠的正當性，接著分類編碼、依據研究目的訂出輸入的特徵及輸出目標。再來第二個困難的點是，在有這麼這麼多的模型可以選擇，到底應該選擇哪個才能最適配資料，又可以讓模型分類、預測的能力最佳，就必須一一熟悉每個模型的運作原理、適用擅長的領域，還透過網格搜索的方法，去找到表現比較好的超參數設定組合，盡量讓模型表現最佳化。然後由於該資料集的失智、非失智人數相差懸殊，在沒有處理資料不平衡的情況之前，模型表現結果可以說是一塌糊塗，為此學習了如何使用SMOTE技術，去平衡資料，也嘗試用class_weight按照資料分布權重去調整模型的損失函數。

最後就是模型解釋，有一大堆的分數指標、線圖，除了要一一算出來數據、畫出每一折的曲線情況、互相比對A、B的結果，還要依據現有的已知資料，嘗試去評價模型的表現，以及解釋造成該結果的原因、特徵影響力、重要度，找出關鍵變項，以利後續能夠在臨床上可以實際應用，協助在早期篩選出病患，以及作為後續治療的依據資料等等。要做完整個報

告，需要學習、具備的資料知識需要非常非常的多，也必須一直重新回頭調整輸入參數、更改模型，不斷不斷重複執行一樣的流程，但也在實際操作的過程中，學習到非常多實際的知識、處理資料的技巧，收穫也還是頗豐富的。

112029024 陳暐凱：

透過這次資料探勘專題，我學會了完整的資料分析流程，包含資料清理、遺漏值處理、特徵選擇與模型建構等實務技能。同時也熟悉了分類與回歸兩種不同任務的分析思維，理解橫斷研究與縱貫研究在研究設計上的差異。此外，藉由 ROC、AUC、R^2 等評估指標的比較與解讀，我對模型效能評估有更清楚的認識，也學會如何透過特徵重要度與結果分析，理解模型背後的判斷依據。整體而言，這次專題加深了我對資料探勘方法與實際應用之間關係的理解。

112029044 蔡碩恩：

這次課程是我在資料科學領域的一次重要實踐。從最初繁瑣的資料清理，到特徵選擇時的取捨，我體會到數據品質對模型結果的決定性影響。我特別受發發的是對評估指標的解讀：ROC 與 AUC 不僅是數字，更是模型在不同閾值下權衡利弊的表現；而對特徵重要度的分析，則讓我學會從數據中挖掘背後的因果邏輯。

理解了橫斷與縱貫研究的設計差異後，我對於如何針對不同維度的數據選擇最適切的探勘方法有了更清晰的藍圖，這對我未來處理複雜研究議題有極大的幫助。