糖尿病預測

一、資料介紹

本研究包含 47,275 筆個體樣本與 23 個變數的資料集，嘗試運用機器學習方法來預測個體是否罹患糖尿病。

本研究資料中包含多項與糖尿病風險相關的重要變數，涵蓋以下幾個面向：

- 生理健康相關變數：是否患有高血壓(HighBP)、是否患有高膽固醇(HighChol)、是否曾中風(Stroke)、是否曾患有心臟病或心臟病發作(HeartDiseaseorAttack)、BMI、是否有走路困難情形(DiffWalk)。
- 健康行為與生活型態變數：是否為吸菸者(Smoker)、是否有在運動 (PhysActivity)、是否每天攝取水果(Fruits)或蔬菜(Veggies)、是否有重度飲酒行為(HvyAlcoholConsump)。
- 醫療可近性與資源使用變數：過去 5 年內是否曾接受過膽固醇檢查(CholCheck)、是否擁有任何醫療保險(AnyHealthcare)、是否曾因費用問題延遲或無法就醫(NoDocbcCost)。
- 自我健康評估與心理狀態變數：整體健康自評(GenHlth)、過去 30 天中精神不佳的天數(MentHlth)、過去 30 天中身體不適的天數(PhysHlth)。
- 人口統計與社經背景變數：性別(Sex)、年齡分組(Age)、教育程度(Education)、收入分級(Income)。

此外，ID 為個體編號，並不會納入模型訓練；Label 為目標變數，用以標示個體是否罹患糖尿病（1 表示罹患，0 表示未罹患）。

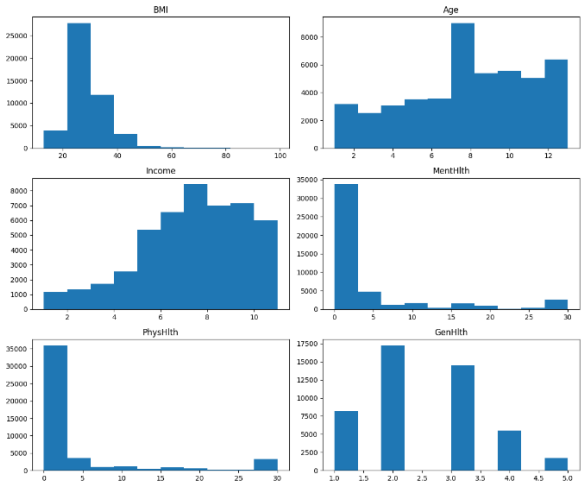
在探索性資料分析中，本研究一律將圖表分為「全體資料」與「Subsample 平衡資料」兩種視角，以避免因資料不平衡造成的分布偏誤。

根據圖一、圖二，Age、Income 在平衡資料後的資料更明顯集中於高齡群與低收入群，推測年齡與糖尿病風險呈正相關，且社經地位可能與糖尿病有關。

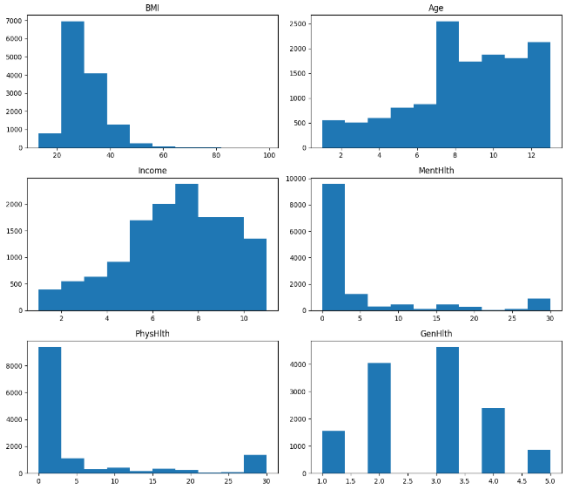
根據圖三、圖四，BMI、MentHlth、PhysHlth 等呈現右偏分布，表示部分個體在這些指標上明顯偏高。Age、BMI、PhysHlth、GenHlth、Income 等變數在是否罹患糖尿病呈現顯著分布差異，糖尿病患者年齡普遍較高，BMI 偏高，身體不適與自評健康狀況較差，且多數收入偏低，這些特徵皆具良好的區辨能力，因此在建模過程中應納入作為重要預測變數。

根據圖九、圖十，HighBP、HighChol、PhysActivity、DiffWalk 在是否罹患糖尿病間具有明顯分布差異。即便在 Subsample 平衡資料下，這些變數仍保有穩定的預測貢獻，尤其以 DiffWalk、HighBP 及 PhysActivity 的差異最為顯著。相對而言，HvyAlcoholConsump、NoDocbcCost、AnyHealthcare 在有無糖尿病之間並無明顯差異，顯示這些變數在本研究中的預測力相對較弱。

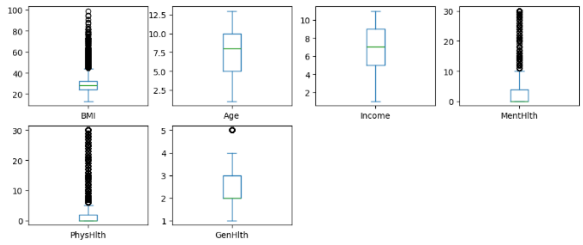
最後，根據圖十二，GenHlth、PhysHlth、DiffWalk 等健康相關變數彼此呈現中度正相關，顯示身體健康不佳的情況往往會同時出現多項症狀。此外，HighBP 與 HighChol 也呈現顯著中度相關，表示高血壓與高膽固醇之間可能存在共病現象。Education 和 Income 也呈現中度正相關，教育程度愈高者，收入通常也較高。但因為這些變數的相關係數不是高度相關，因此本研究在建模時不會先刪除變數。



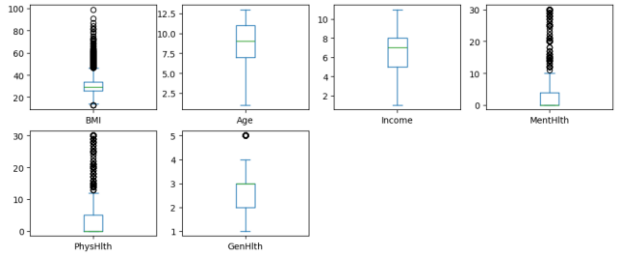
圖一、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 長條圖(所有資料)



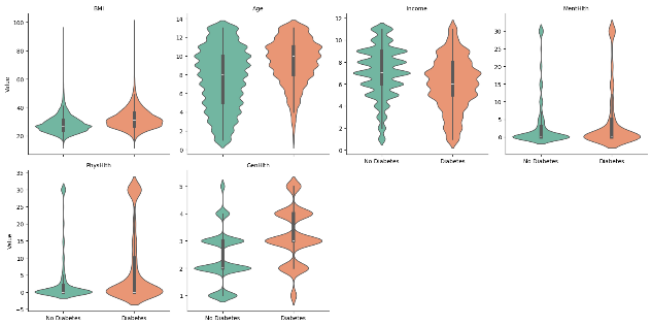
圖二、MI、Age、Income、MentHlth、PhysHlth、GenHlth 長條圖(Subsample)



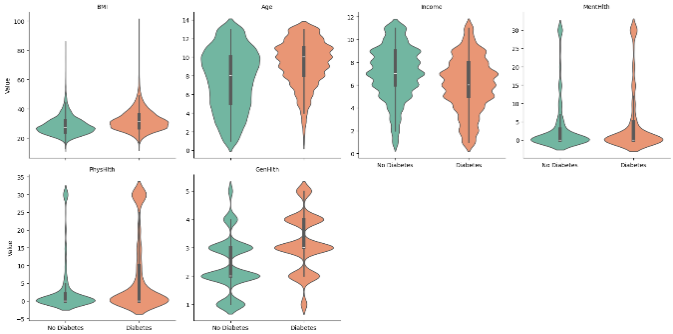
圖三、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 的箱型圖(所有資料)



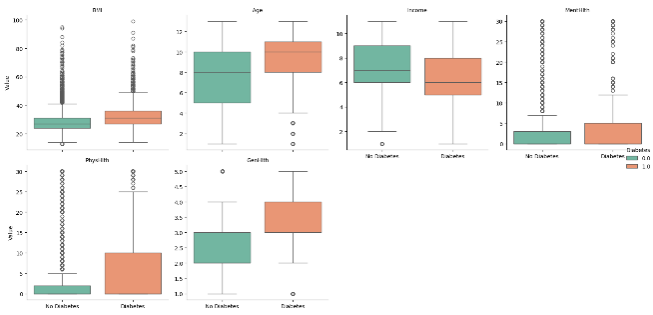
圖四、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 的箱型圖(Subsample)



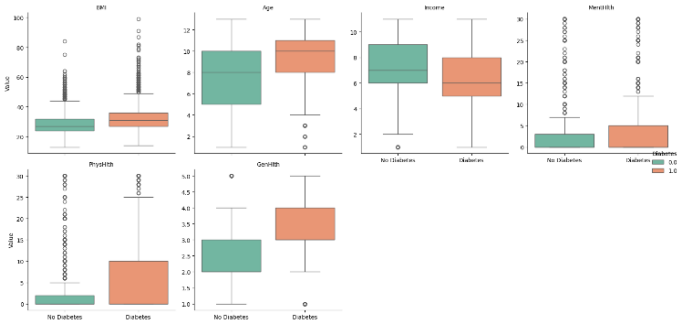
圖五、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 在糖尿病與非糖尿病族群中的 vilolin plot (所有資料)



圖六、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 在糖尿病與非糖尿病族群中的 vilolin plot (subsample)

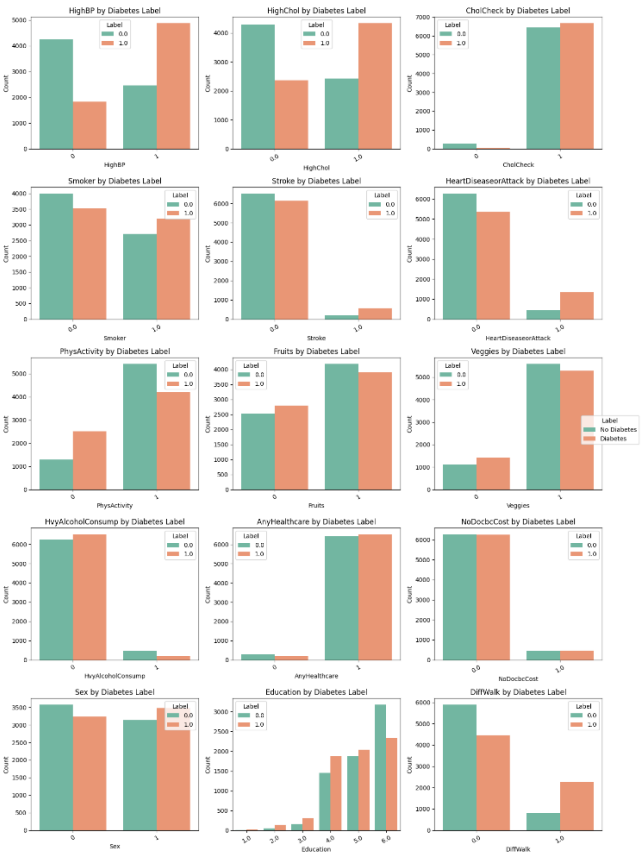


圖七、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 在糖尿病與非糖尿病族群中的箱型圖(所有資料)

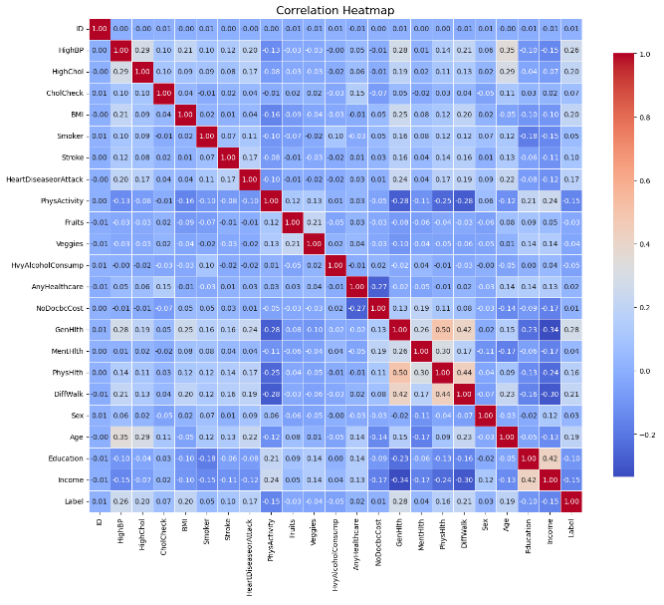


圖八、BMI、Age、Income、MentHlth、PhysHlth、GenHlth 在糖尿病與非糖尿病族群中的箱型圖(subsample)

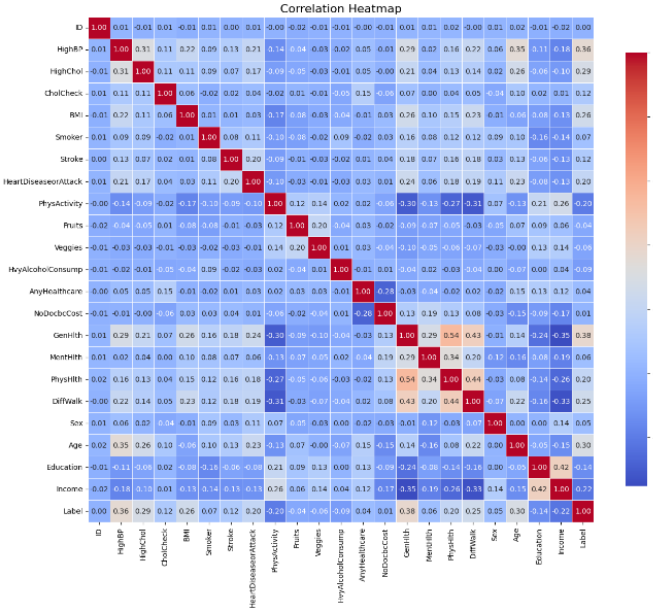
圖九、HighBP、HighChol、CholCheck、Smoker、Stroke、HeartDiseaseAttack、PhysActivity、Fruits、Veggies、HvyAlcoholConsump、AnyHealthcare、NoDocbcCost、 Sex、Education、DiffWalk 在糖尿病與非糖尿病族群中的長條圖(所有資料)



圖十、HighBP、HighChol、CholCheck、Smoker、Stroke、HeartDiseaseAttack、PhysActivity、Fruits、Veggies、HvyAlcoholConsump、AnyHealthcare、NoDocbcCost、Sex、Education、DiffWalk 在糖尿病與非糖尿病族群中的長條圖(subsample)



圖十一、所有資料 Heatmap



圖十二、Subsample 資料 Heatmap

二、模型實驗

本研究比較以下幾種方法：

1. SMOTE → Random Forest + Random Search
2. SMOTE → XGBoost + Random Search
3. SMOTE → CatBoost + Random Search
4. SMOTE → LightGBM + Random Search
5. SMOTE → Stacking（XGBoost / CatBoost / LightGBM）
6. Feature Selection → SMOTE → XGBoost + Random Search
7. Feature Selection → SMOTE → CatBoost + Random Search
8. Standardization → SMOTE → MLP（神經網路）

針對不平衡資料問題，本研究會使用 SMOTE 來進行類別平衡化處理。模型分別會分別使用 Bagging 的 Random Forest 模型、Boosting 的 Xgboost、Catboost、LightGBM。所有模型會透過 Random Search 來找最合適的參數，並且用 macro F1-score 為評分指標，對訓練資料做 5 fold Cross Validation 來驗證模型穩定性。由於原始資料中樣本分布極度不均，因此我加入調整預測機率 threshold 的策略，藉此提升模型對少數類別的辨識能力，使模型輸出更符合實際需求。在調整過程中測試了 0.2 至 0.5 之間的 threshold，最終發現 threshold 設為 0.3 時能達到最佳的 macro F1-score，因此本研究採用 0.3 作為最終預測門檻值。

從表一結果可以觀察到，XGBoost 與 CatBoost 在驗證集與 Public Leaderboard 上的表現皆優於 LGBM 與 Random Forest，顯示 Boosting 模型較能處理此類表格型、非均衡資料。而 Random Forest 表現相對較差，所以未進一步在 Public Leaderboard 進行測試。

進一步探討模型結合方式時，我們嘗試將三種 Boosting 模型進行 Stacking，期望能提升預測能力，但結果反而略遜於單一模型，所以我決定嘗試用特徵選擇的方法來強化模型：

- 在 XGBoost 中，根據圖十三和先前的 EDA 結果，決定移除 AnyHealthcare 變數。儘管 MentHlth 、 PhysHlth 在 XGBoost 中的重要性偏低，但 EDA 顯示在糖尿病患者與非患者間具有顯著分布差異，顯示仍具一定辨識力。進一步分析發現，MentHlth 與 PhysHlth 與 GenHlth、DiffWalk 呈中度正相關，推測其預測能力可能被其他相關變數所分擔，進而降低在模型中的個別重要性。基於其間接貢獻，本研究選擇保留這兩個變數。經刪除變數 AnyHealthcare 後，Macro F1 分數由 0.6707 提升至 0.6711。
- 在 CatBoost 中，根據圖十四，我們刪除了 CholCheck、Stroke、AnyHealthcare 三個重要性最低的變數。但此策略 Macro F1-score 由 0.6710 略降至 0.6708。

此外，我們也嘗試 MLP 神經網路應用，先將資料進行標準化，再使用 SMOTE 處理不平衡問題。模型設計包含兩層隱藏層，搭配 BatchNorm、Dropout 與 ReLU 激活函數，並以 Sigmoid 輸出機率。儘管訓練損失持續下降，驗證集上的 Macro F1-score 卻隨訓練進行而下降，呈現出過擬合與泛化能力不足的現象，顯示 MLP 不適合處理此類表格型結構資料。

根據表一，進行特徵選擇後的 XGBoost 在驗證集和 Public 測試資料上表現最佳，最後會選擇 Feature Selection → SMOTE → XGBoost 作為最後提交的版本。

	Cross Validation 下的 F1-score	Public 50%
SMOTE→Random Forest	0.6456	效果太差沒有試
SMOTE→Xgboost	0.6707	0.67202
SMOTE→Catboost	0.6710	0.67345
SMOTE→LGBM	0.6697	0.67390
SMOTE→Stack (Xgboost/ Catboost/ LGBM)	0.6622	0.67017
Feature selection→SMOTE→Xgboost	**0.6711**	**0.67779**
Feature selection→SMOTE→Catboost	0.6708	0.67269
Standardize→SMOTE→神經網路	0.56~0.58	效果太差沒有試

表一、不同方法驗證集和 50% public test data F1-score

	Feature	Importance
0	HighBP	624.406860
1	HighChol	434.002472
7	PhysActivity	234.202240
8	Fruits	217.421173
4	Smoker	213.216614
17	Sex	200.822311
13	GenHlth	188.217880
6	HeartDiseaseorAttack	148.280762
16	DiffWalk	141.817078
10	HvyAlcoholConsump	103.972702
9	Veggies	87.536537
19	Education	75.871773
18	Age	67.681854
3	BMI	45.118835
20	Income	36.636871
2	CholCheck	35.240078
12	NoDocbcCost	34.239597
5	Stroke	30.996819
14	MentHlth	24.273014
15	PhysHlth	17.044991
11	AnyHealthcare	6.305292

圖十三、Xgboost 的特徵重要性排序

	Feature	Importance
1	HighChol	11.691051
17	Sex	11.371342
8	Fruits	9.795424
4	Smoker	9.736527
13	GenHlth	9.390140
0	HighBP	8.418592
19	Education	7.564058
7	PhysActivity	7.508541
9	Veggies	5.218226
18	Age	4.690911
20	Income	4.350391
6	HeartDiseaseorAttack	3.741619
16	DiffWalk	3.319712
3	BMI	1.028298
14	MentHlth	0.681178
15	PhysHlth	0.667240
12	NoDocbcCost	0.343123
10	HvyAlcoholConsump	0.321158
2	CholCheck	0.116429
5	Stroke	0.043991
11	AnyHealthcare	0.002047

圖十四、Catboost 的特徵重要性排序

三、討論

1. 結果探討

a. Random forest 為什麼表現比較差

Random Forest 是 Bagging 模型，擅長處理高變異資料與避免過擬合，但在高度不平衡的分類問題中，對少數類別的辨識能力有限。相較之下，Boosting 模型（如 XGBoost、CatBoost、LGBM）能針對錯誤分類樣本進行強化學習，更能聚焦於難以分類的少數類別，因此在 Macro F1-score 指標上表現更佳。

b. 使用 Stack model 為什麼表現比較差

原先預期不同模型具有互補性，將預測結果進行 Stacking 可望提升泛化能力。觀察 XGboost、CatBoost、LGBM 模型的相關係數，XGboost 和 CatBoost 的相關係數為 0.9876、XGboost 和 LGBM 為 0.9943 和 CatBoost 和 LGBM 為 0.9814，顯示三者預測結果高度相似。由於缺乏多樣性，Stacking 並未帶來顯著的整體效能提升，反而因增加模型複雜度與過擬合風險，而導致 Macro F1-score 微幅下降。

c. 使用神經網路為什麼表現比較差

儘管神經網路在影像與語音等非結構化資料上表現卓越，但對於結構化、表格型資料，優勢相對有限。這可能是因為資料變數規模不大，特徵之間邏輯結構更適合傳統機器學習模型，而非需要大量參數調整與大樣本訓練的神經網路。這也再次印證了 Boosting 在表格型資料任務中的優勢。

2. 未來展望

a. 本研究可進一步嘗試多元的特徵工程技巧，例如加入特徵交互作用，以凸顯與糖尿病高度相關的變數組合，進一步提升模型預測表現。

b. 儘管本研究 XGBoost 表現穩定，但與 CatBoost、LGBM 等 Boosting 模型高度相關，缺乏互補性。未來可考慮加入如 SVM、Logistic Regression、KNN 等，並使用 VotingClassifier 或 StackingClassifier 等集成方法，結合多模型優勢，有機會進一步提升模型預測表現。