利用健康數據資料和機器學習技術預測糖尿病發生機率

張哲安

一、研究背景與動機

根據 Petersmann et al. (2019) 提出對糖尿病的定義為「一種代謝症候群，其中一項為慢性糖尿病，導致原因為胰島素分泌不足或無法分泌或兩者的結合。」，造成的原因有可能為「肥胖、缺乏運動、藥物或疾病、年齡或家族遺傳」[1]。當身體處於長時間的高血糖，造成身體許多器官的損害，常見合併症包含「高血壓、心臟疾病、腎臟病、視網膜病變……等。」[2]。根據衛生福利部國民健康署指出「全國約有 200 多萬名糖尿病患者，每年以 25,000 名的速度持續增加，糖尿病為國人十大死因之一，近萬人因為糖尿病及引發的併發症而死亡。」[3]，根據世界衛生組織統計，約有 70% 的人口是屬於疾病尚未發生，但處於罹患疾病的風險當中[4]。Joslin (2021) 表示「糖尿病患者常會在疾病晚期才發現及接受治療，如果能夠早期發現，更容易通過飲食來治療。」，透過早期改變良好生活習慣及定期追蹤，控制血糖，延緩及避免併發症的發生。所以能夠有效精準的預測糖尿病發生機率為重要議題。

近年來人工智慧 (artificial intelligence) 技術的進步，能夠根據大數據來預測疾病的發生風險，即可提早進行預防、就醫或治療，以避免疾病之發生或減少疾病發生所造成之傷亡。故本文通過機器學習 (machine learning) 方式包含決策樹、XGBoost[5] 及自動化機器學習 (Automatic Machine Learning；以下簡稱 AutoML) [6] 預測糖尿病的發生機率。

二、研究資料分析

本文資料來源為 Kaggle 網站上取得 Diabetes Health Indicators Dataset[7]，其資料樣本 (sample) 來源於美國 CDC 使用電話調查取得健康相關資訊。總計有 70,693 筆資料，其中以有無糖尿病做為此資料的標籤 (labels)，也就是想要預測的目標變數 (target)，兩者樣本數皆為 35,346 筆，為平衡資料 (balanced data)。此外，每筆資料均包含 21 項健康相關資訊特徵 (features)，如：有無高血壓、有無高血脂、有無血脂檢查、BMI 身體質量指數、有無抽菸、有無中風、有無心臟疾病、有無運動、有無吃水果、有無吃蔬菜、有無重度飲酒、有無醫療保險、有無因沒錢無法就醫、一般身體健康程度、心理健康程度、生理健康程度、有無困難走路、性別、年齡、學歷及收入，詳細特徵定義詳見附錄。

三、建立預測模型與分析

本文希望利用健康相關資訊之 21 項特徵資料進行預測糖尿病的發生機率，故選擇使用

[1] https://www.chp.gov.hk/tc/static/80037.html
[2] https://www.commonhealth.com.tw/article/83965
[3] https://www.hpa.gov.tw/Pages/List.aspx?nodeid=359
[4] http://www.fma.org.tw/2018/bio-3.html
[5] https://xgboost.readthedocs.io/en/stable/
[6] https://learn.microsoft.com/zh-tw/azure/machine-learning/concept-automated-ml?view=azureml-api-2
[7] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

監督式學習 (supervised learning)[8] 中分類演算法 (classification algorithm)，包含決策樹 (decision Tree)[9]、Chen (2016) 提出的 XGBoost 及 AutoML。其中決策樹是利用特徵資料不斷分類分支有無糖尿病，最後形成似一棵樹。XGBoost 可視為產生多棵決策樹，而每一棵新的決策樹都會優化、修正前一棵決策樹預測錯誤的資料。無論是決策樹或 XGBoost 都使用交叉驗證 (Cross Validation, CV) 得出模型 (model) 的超參數 (hyperparameter)。本文使用 AutoML 中的 PyCaret[10] 套件，此套件可以利用內建設定好每個模型超參數範圍，自動利用交叉驗證得出每個模型最佳的超參數來訓練模型，再自動進行每個模型結果的比較，最後得出一個最適合此資料的最佳模型及其對應最佳超參數。

為使模型擁有預測未知資料的能力，因此將資料隨機 (random) 分割成訓練資料 (train set) 及測試資料 (test set) 為 7：3，運用訓練資料來訓練並建立預測模型後，將測試資料視為未來未知需要預測的資料，來評估此模型的是否適合用於預測未來資料。

依據網路資源：「大部分使用內建超參數建立出的模型，容易出現過度配適 (overfitting) 的現象，而進行調參數 (hyperparameters tuning) 是可以解決此問題的其中一種方式。」[11]，其中過度配適是指模型在訓練資料上可以有很好的預測結果，但在測試資料上無法很好的預測結果，且兩者間出現明顯差異。常用調參數的方法為交叉驗證，「最常見的為 K-fold，即假設 K=5，也就是將訓練資料切割為 5 等份，相同的模型需要訓練 5 次，每次訓練都會從這 5 等份中挑選 4 等份作為訓練資料，剩 1 等份沒有參與訓練的資料作為驗證資料 (validation set)，並且每一等份資料都會被作為驗證資料。」[12]。

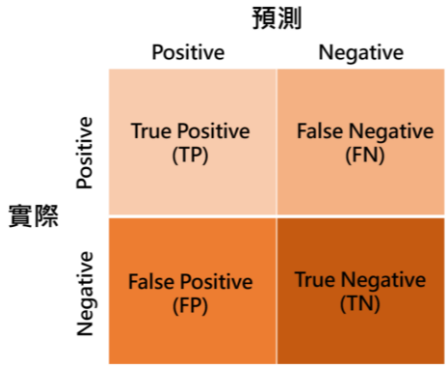
文獻上利用交叉驗證調參數最常見的兩種方式為：網格搜尋交叉驗證 (GridSearchCV)[13] 及隨機搜尋交叉驗證 (RandomizedSearchCV)[14]。依據網路資源：「隨機搜尋交叉驗證透過一定範圍內隨機的數值組合，可以彌補網格搜尋交叉驗證固定有限組合之缺點，更可以找出最佳超參數組合。」[15]。雖然於決策樹、XGBoost、PyCaret 都可以使用其他搜尋方式得到最佳超參數，但在本文中是使用隨機搜尋交叉驗證得出最佳超參數下，比較決策樹、XGBoost、PyCaret 模型的預測結果。

在比較不同模型預測有無糖尿病或糖尿病發生機率結果上，使用三項分類預測常見的評估指標，分別為 AUC score、F1 score、Accuracy score。因標籤為實際上有無糖尿病兩種結果（有標記為正、無標記為負），預測結果有預測正確及錯誤兩種結果（正確標記為正、錯誤標記為負），故共有四項組合列於表一，其四種狀況分別為「TP (True Positive)、TN (True Negative)、FP (False Positive)、FN (False Negative)。」[16]，分析預測結果時常以混淆矩陣 (confusion matrix) 顯示如圖一。

[8] https://ithelp.ithome.com.tw/articles/10196922
[9] https://scikit-learn.org/stable/modules/tree.html
[10] https://pycaret.org/
[11] https://axk51013.medium.com/ml 調參數神物-successive-halving-sklearn-0-24 開始支援-680f05c56519
[12] https://jason-chen-1992.weebly.com/home/-cross-validation
[13] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
[15] https://blog.csdn.net/BF02jgtRS00XKtCx/article/details/112975337
[16] https://blog.csdn.net/BF02jgtRS00XKtCx/article/details/112975337

表一：分類模型預測結果的四種可能

英文全文	縮寫	實際與預測狀況	預測正確與否
True Positive	TP	實際為正，且預測為正。	預測正確
False Negative	FN	實際為正，但預測為負。	預測錯誤
True Negative	TN	實際為負，且預測為負。	預測正確
False Positive	FP	實際為負，但預測為正。	預測錯誤



圖一：混淆矩陣 (圖片來源：https://medium.com/marketingdatascience/分類器評估方法-roc 曲線-auc-accuracy-pr 曲線-d3a39977022c)

本文將採用以下三項分類預測評估指標為：

1. AUC (Area Under Curve) score 是「ROC 曲線下的面積，ROC 曲線是由 X 軸的偽陽性率 (false positive rate) 及 Y 軸的真陽性率 (true positive rate) 畫製而成。偽陽性率為預測是正確，但實際上是錯誤，真陽性率為預測是正確的，實際也是正確。」[17]。依據網路資源：「AUC 代表意義為正確判斷為正樣本的機率高於錯誤判斷為負樣本的機率。」[18]。AUC 也可以說是真陽性率的平均。
2. F1 score 是「召回率 (recall)、精確率 (precision) 兩者指標的整合。召回率是所有實際為正確數量當中 (TP+FN)，能夠預測多少正確 (TP) 的比例，準確率為在所有預測為正確的中 (TP+FP)，有多少為預測正確 (TP) 的比例。」[19]，其數學公式為：

$$recall = \frac{TP}{TP + FN},$$

$$precision = \frac{TP}{TP + FP},$$

$$F1\ score = \frac{2}{\frac{1}{recall} + \frac{1}{precison}}.$$

17 https://www.cupoy.com/qa/club/ai_tw/0000016D6BA22D97000000016375706F795F72656C656173654B5741535354434C5542/0000017B7B5B4859000000246375706F795F72656C656173655155455354

18 https://tsupei.github.io/nlp/2019/12/30/prcurve.html#:~:text=AUC (曲線下面積)，所代表的意義為隨機抽取一個正樣本，分類器會正確判斷為正樣本的機率高於誤判斷為負樣本的機率，所以 AUC 越高則分類器正確率會越高。

19 https://medium.com/nlp-tsupei/precision-recall-f1-score 簡單介紹-f87baa82a47

3. Accuracy score 是「準確率，是所有預測正確的數量 (TP+TN) 佔整體 (TP+TN+FP+FN) 的比例。」[20]。

此三項評估指標數值均介於 0 與 1 之間，且數值越大代表模型的預測效果越好。

以下是使用隨機搜尋交叉驗證得出最佳超參數下，比較決策樹、XGBoost、PyCaret 的模型結果。為比較之公平性，相同的超參數設定為一致。設定隨機種子 (random seed) 為 42，保證重複執行程式有相同結果。設定隨機抽取組合數為 500，總共進行 500 種超參數組合。設定模型超參數組合評估指標及比較模型評估指標為 AUC，表示使用 AUC 評分方式決定最佳超參數組合及最佳模型的比較。設定交叉驗證分割數為 5。

(一) 決策樹

在使用決策樹模型時，隨機搜尋交叉驗證中超參數的選擇及範圍，參考資料來源於網路資料[21]及 Bing 的 GPT-4，設定五項超參數，分別為 criterion、max_depth、splitter、max_features、min_samples_leaf，表二為詳細超參數定義及設定範圍。

表二：決策樹使用隨機搜尋交叉驗證超參數設定

超參數	超參數挑選範圍	超參數意義
criterion	Entropy、Gini	熵 (entropy)、gini 不純度 (gini impurity) 兩種選項，透過不同算法找出最佳節點及分枝的指標。
max_depth	1~15	決策樹的深度或層數。
splitter	Best、Random	透過隨機或優先選擇更重要的特徵進行決策樹的分枝。
max_features	1~21	在進行決策樹的分支時需考慮多少個特徵，因本資料特徵共 21 項故設定為 1 至 21。
min_samples_leaf	50~100	在每次分支完至少有多少資料才進行分支。

表三為使用隨機搜尋交叉驗證超參數下的決策樹的評估指標結果，最佳超參數為：criterion 為 gini、max_depth 為 9、splitter 為 best、max_features 為 14、min_samples_leaf 為 72。從表三可發現在訓練、測試資料 AUC score、F1 score、Accuracy score 皆差異不大，所建立的決策樹模型並沒有過度配適的現象。

表三：使用隨機搜尋交叉驗證超參數下的決策樹各項評估指標分數

AUC score		F1 score		Accuracy score	
訓練資料	測試資料	訓練資料	測試資料	訓練資料	測試資料
0.8295	0.8174	0.7629	0.7543	0.7523	0.7420

(二) XGBoost

在使用 XGBoost 模型時，隨機搜尋交叉驗證中超參數的選擇及範圍，超參數的選擇來源

[20] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html#sklearn.metrics.auc
[21] https://blog.csdn.net/u010591976/article/details/105825363

於網路參考[22]及 Bing 的 GPT-4，設定四項超參數，分別為 n_estimators、learning_rate、max_depth、colsample_bytree，表四為詳細超參數定義及設定範圍。

表四：XGBoost 使用隨機搜尋交叉驗證超參數設定

超參數	超參數挑選範圍	超參數意義
n_estimators	100~500	總共會產生出多少棵決策樹，此處設定為 100 至 500 棵。
learning_rate	0.01~0.3(每次間隔 0.01)	控制每一棵新的決策樹對最終預測結果的影響，主要用於防止過度配適。
max_depth	1~15	決策樹的深度或層數，和決策樹設定相同。
colsample_bytree	0.1~1(每次間隔 0.1)	控制每棵決策樹使用多少項特徵，其目的是防止過度配適。

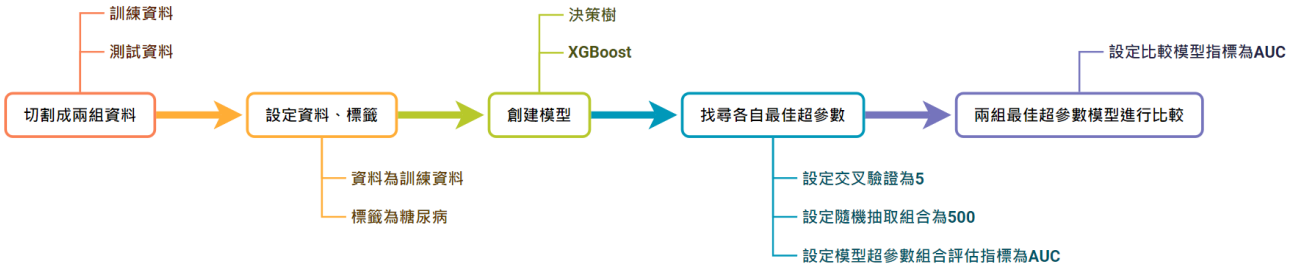
表五為使用隨機搜尋交叉驗證超參數下的 XGBoost 的評估指標結果，最佳超參數為：n_estimators 為 271、learning_rate 為 0.05、max_depth 為 5、colsample_bytree 為 0.3。從表五可發現在訓練、測試資料 AUC score、F1 score、Accuracy score 皆差異不大，所建立的 XGBoost 模型並沒有過度配適的現象。

表五：使用隨機搜尋交叉驗證超參數下的 XGBoost 各項評估指標分數

AUC score		F1 score		Accuracy score	
訓練資料	測試資料	訓練資料	測試資料	訓練資料	測試資料
0.8427	0.8332	0.7708	0.7675	0.7614	0.7571

（三）AutoML PyCaret

本文使用 AutoML 其中一個套件 PyCaret，圖二為其做法步驟流程圖。因本文比較決策樹及 XGBoost，故由 PyCaret 自動創建決策樹及 XGBoost 兩種模型。值得注意的是之前的隨機搜尋交叉驗證是利用網路資料及 Bing 的 GPT-4，設定決策樹共 5 項超參數及 XGBoost 共 4 項超參數，而 PyCaret 不一樣之處為自動使用內建設定好的決策樹共 6 項超參數及 XGBoost 共 9 項超參數進行隨機搜尋交叉驗證，並可以自動比較多個模型。



圖二：使用 PyCaret 流程圖

從表六可發現最佳模型為 XGBoost，在訓練、測試資料 AUC score、F1 score、Accuracy

[22] https://stackoverflow.com/questions/69786993/tuning-xgboost-hyperparameters-with-randomizedsearchcv

score 皆差異不大，所建立的 XGBoost 模型並沒有過度配適的現象。雖然於 AUC score 中測試資料評估指標大於訓練資料 0.000002，但可以說差異非常小。

表六：使用 AutoML PyCaret 得出最佳模型及超參數模型

最佳模型	AUC score		F1 score		Accuracy score	
XGboost	訓練資料	測試資料	訓練資料	測試資料	訓練資料	測試資料
	0.832320	0.832322	0.7456	0.7451	0.7500	0.7485

（四）決策樹、XGBoost、PyCaret 模型比較

為方便比較不同模型結果，將表四、五、六整合為表七，並將最佳結果以顏色顯示出來。表七為決策樹、XGBoost、PyCaret 模型的 AUC score、F1 score、Accuracy score 訓練、測試資料評估指標，三個模型的三項評估指標，並沒有過度配適的現象，以下為三個模型比較結果：

1. 決策樹與 XGBoost 比較：在 XGBoost 中，三項評估指標在訓練及測試資料上，都較優於決策樹，也驗證本文之前所說 XGBoost 會優化、修正前一棵決策樹預測錯誤的資料，擁有更好的預測能力。
2. 決策樹與 PyCaret 比較：F1 score 的訓練、測試資料、Accuracy score 的訓練資料，PyCaret 都低於決策樹，但差異較小。
3. XGBoost 與 PyCaret 比較：PyCaret 三項評估指標在測試及訓練資料皆較 XGBoost 差。

由於 PyCaret 的結果包含更完整的調參數及不同模型間的比較，直覺上 PyCaret 三項評估指標應要優於決策樹及 XGBoost，但這和上述比較結果 2 和 3 並不一致。經反思研究，可能原因為 PyCaret 的決策樹需要在 6 項超參數中找出最佳超參數，PyCaret 的 XGBoost 需要在 9 項超參數中找出最佳超參數，較本文設定決策樹共 5 項超參數及 XGBoost 共 4 項超參數多，且 PyCaret 之每項超參數設定的範圍更大，組合數為更多，可能需要更多次的隨機抽取組合 (n_iter)，才能對這組分析資料取得更好的最佳超參數組合。

表七：比較決策樹、XGBoost、PyCaret 模型

模型	AUC score		F1 score		Accuracy score	
	訓練資料	測試資料	訓練資料	測試資料	訓練資料	測試資料
決策樹	0.8295	0.8174	0.7629	0.7543	0.7523	0.7420
XGBoost	0.8427	0.8332	0.7708	0.7675	0.7614	0.7571
PyCaret	0.8323	0.8323	0.7456	0.7451	0.7500	0.7485

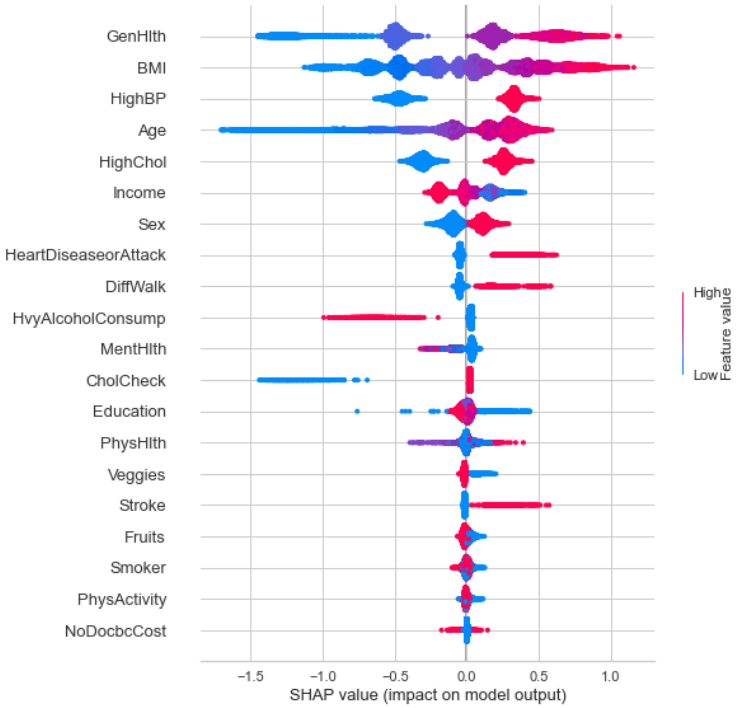
四、可解釋人工智能

依據網路資源：「當在使用較簡單模型如決策樹模型時，可以由所建立的預測模型了解預測結果與特徵之間的關係。但當使用較複雜的模型如 XGBoost 模型時，就無法輕易由所建立的預測模型了解預測結果與特徵之間的關係。」[23]，因此當使用較複雜的模型時，就不容易對

[23] https://medium.com/@jimmywu0621/可解釋ａｉ-什麼是 shap-5ec3953e3c5b

預測結果進行合理解釋，這在實務應用上就容易造成困擾。而可解釋人工智能(Explainable Artificial Intelligence；以下簡稱可解釋 AI)是一種可以了解預測結果之原因的方法。另外依據網路資源：「2018 年歐盟修訂的一般資料保護規範 (General Data Protection Regulation) 宣布要求解釋的權力，希望以此解決近來備受重視的演算法可能帶來的道德與社會法律問題。」[24]。在可解釋 AI 方法中，常利用 Shapley (1951) 提出的夏普利值 (Shapley Value；以下簡稱 SHAP value) 的相關做法，其優點為可做全域解釋性 (global interpretation)，也能達成局部解釋性 (local interpretation)。全域解釋性為整體特徵如何影響預測結果，局部解釋性為每筆資料特徵如何影響預測結果。有關可解釋 AI 進一步介紹內容可參考 Masís (2021) 及 Molnar (2022)。由第三節可得知這組資料對 XGBoost 模型有較好的結果，故本文使用 XGBoost 模型以訓練資料進行可解釋 AI，其結果為圖三。

圖三中之每個點代表每筆資料對應每項特徵其影響預測結果的貢獻值，稱為 SHAP value。當 SHAP value 為零時，代表此特徵對預測結果無影響。當 SHAP value 大於零，代表此特徵對預測結果有著正向影響。當 SHAP value 小於零時，代表此特徵對預測結果有著反向影響，另外當 SHAP value 正數值越大或負數值越小，影響預測結果越大。圖三中之左側為影響是否有糖尿病的各項特徵，由上至下依序代表特徵影響預測結果由大到小。圖三中之顏色每筆資料對應特徵數值大到小，顏色的呈現為紅到藍。當特徵內容為 0 或 1 兩類時，顏色分別以藍或紅表示，但當特徵內容非屬 0 或 1 兩類的多個數值時，顏色是依數值由小至大，依序由藍色逐漸變為紅色。



圖三：可解釋 AI 圖

圖三中可發現對整體資料而言 (全域解釋) 影響預測糖尿病發生機率的前五名特徵為 1.自覺健康評分 2.身體質量指數 (BMI) 數值 3.高血壓 4.年齡 5.高血脂。而最重要的特徵為自覺健康評分，此項特徵定義中，分數越高代表覺得自身健康較差，對應顏色為紅色，而分數

[24] https://zh.m.wikipedia.org/zh-tw/可解釋人工智慧

越低則代表覺得自身健康較好，對應顏色為藍色。從圖三可看到紅色向右延伸，SHAP value 大於零，代表當自評身體健康分數越高、身體越差時，預測糖尿病發生機率較大。反之，藍色向左延伸，代表當自評身體健康分數越低、身體越好時，SHAP value 小於零，預測糖尿病發生機率較小。其可能原因為糖尿病大致可分為第一型、第二型糖尿病及其他因素引起的糖尿病。第一型通常在出生時因自體免疫失調摧毀自身胰島素導致，而第二型通常認為與生活習慣及肥胖有關。另外也會因外傷、胰臟發炎、基因異常、自體免疫疾病、因疾病需要長期服用類固醇等原因引起的糖尿病。無論如何，由於上述原因可能是糖尿病的致病原因，所以在自覺健康上評分往往都較差。

影響預測糖尿病發生機率的第二項特徵為身體質量指數 (BMI) 數值，當數值越大顏色為紅色向右延伸，SHAP value 大於零，預測糖尿病發生機率越大。反之當數值越小顏色為藍色向左延伸，SHAP value 小於零，預測糖尿病發生機率越小。其可能原因為當體重過重時身體會製造更多的胰島素來控制血糖，但是當胰臟無法分泌足夠的胰島素時，造成血糖失控，就容易引起糖尿病。

影響預測糖尿病發生機率的第三項特徵為高血壓，患有高血壓者顏色為紅色向右延伸，SHAP value 大於零，預測糖尿病發生機率大。反之沒有患有高血壓顏色為藍色向左延伸，SHAP value 小於零，預測糖尿病發生機率越小。其可能原因為患有高血壓者致病原因多為飲食因素，其中一項為攝取過量反式脂肪，而含有反式脂肪最常見的食品為奶油，多用於麵包、蛋糕、糕點類食品。然而在這些食品中都常含有精緻糖，當精緻糖進入身體後，血糖會快速上升，啟動胰島素作用，就導致血糖快速下降，而當低血糖時，會想更快且攝取更多的糖份，就造成血糖控制不佳，較容易引起糖尿病。

影響預測糖尿病發生機率的第四項特徵為年齡，當年齡越大顏色為紅色向右延伸，SHAP value 大於零，預測糖尿病發生機率越大，當年齡越小顏色為藍色向左延伸，SHAP value 小於零，預測糖尿病發生機率越小。其可能原因為隨著年齡的增長、器官的老化，其中胰臟退化，使得功能減退，胰島素分泌減少，而導致血糖的控制不穩，較容易引起糖尿病。

影響預測糖尿病發生機率的最後一項特徵為高血脂，當患有高血脂顏色為紅色向右延伸，SHAP value 大於零，預測糖尿病發生機率越大，反之沒有罹患高血脂顏色為藍色向左延伸，SHAP value 小於零，預測糖尿病發生機率小。其可能原因為當血脂過高時，會導致胰島素抗性，亦即身上各種細胞對所分泌的胰島素無反應，而需要更高胰島素濃度，導致血糖忽高忽低，影響血糖的控制，較容易引起糖尿病。

我們所熟知的三高為高血糖、高血脂、高血壓，而導致三高的危險因子中有各式疾病、年齡及肥胖，此六項形成複雜循環的閉鎖網絡，環環相扣互相影響。高血糖症狀通常意味著患有糖尿病，因此高血糖不在本資料特徵中，為本資料中的標籤，剩下五項正是可解釋 AI 中影響預測為糖尿病機率的前五項。

此外，HvyAlcoholConsump (重度飲酒)特徵之 SHAP value 結果與直覺似乎不一致，有重度飲酒者顏色為紅色向左延伸，SHAP value 小於零，預測糖尿病發生機率越小，沒有重度飲酒者顏色為藍色無向外延伸，SHAP value 等於零，不影響糖尿病發生機率。這代表喝酒可以減少糖尿病發生機率，而不喝酒不會對糖尿病有任何影響，這和直覺不一致。但是根據 Pietraszek et al. (2010) 指出「長期且平均的酒精攝取可以改善血糖控制，輕度至中度的酒精攝取可降低 30% 糖尿病的風險，可能與胰島素抗性有關，並且還有可以降低三酸甘油脂及低

密度膽固醇，增加高密度膽固醇的效果。」。此處胰島素抗性是指酒精可以改善各種細胞對原本所分泌的胰島素無反應，需要更高胰島素濃度的問題，導致血糖忽高忽低而罹患糖尿病。同時輕度至中度的酒精攝取也會降低罹患高血脂或肥胖導致 BMI 高的可能性，進而降低罹患糖尿病。雖然酒精可以降低糖尿病、高血脂及肥胖導致 BMI 高等問題，但 Pietraszek et al. (2010) 亦指出「需注意的是有可能發展成酒精依賴成癮的問題，及肝炎、胰臟炎……等疾病，酒精帶來的有害影響可能超過帶來的有益影響。」。

五、結論與未來研究方向

在預測糖尿病發生機率上，使用美國 CDC 透過電話調查取得的 21 項健康相關資訊，利用訓練資料建立預測糖尿病發生機率模型，可以得到 XGBoost 為最佳預測糖尿病發生機率模型，並且在訓練、測試資料，皆沒有過度配適的現象。另外由可解釋 AI 可知對整體資料而言，影響預測糖尿病發生機率的前五名依序為自覺健康評分、身體質量指數 (BMI) 數值、高血壓、年齡及高血脂，且此五項皆為學理上糖尿病發生的常見原因。

未來值得研究方向如下:

1. 除了本文中使用的決策樹、XGBoost，可進一步使用隨機森林 (random forest)、CatBoost、LightGBM 及類神經網路 (neural network) 等方法。
2. 自動化機器學習 PyCaret 的其他應用。
3. 可解釋 AI 之 SHAP value 其他方法的應用。
4. 可解釋 AI 方法亦可以使用由 Ribeiro et al. (2016) 提出局部可解釋模型 LIME (Local Interpretable Model-Agnostic Explanations) 的應用。
5. 對此資料中特徵進行公平性 (fairness) 的探討，了解特徵對預測結果是否有偏見或歧視，導致影響預測結果。

參考文獻

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Joslin, E. P. (2021). The Prevention of Diabetes Mellitus. *Jama*, *325*(2), 190.

Masís, S. (2021). *Interpretable Machine Learning with Python: Learn to Build Interpretable High-performance Models with Hands-on real-world Examples*, Packt Publishing.

Molnar, C. (2022). *Interpretable Machine Learning:A Guide for Making Black Box Models Explainable (2nd Ed.)*, Independently Published.

Petersmann, A., Müller-Wieland, D., Müller, U. A., Landgraf, R., Nauck, M., Freckmann, G., Heinemann, L., & Schleicher, E. (2019). Definition, Classification and Diagnosis of Diabetes Mellitus. *Experimental and Clinical Endocrinology & Diabetes, 127*(1), 1–7.

Pietraszek, A., Gregersen, S., & Hermansen, K. (2010). Alcohol and type 2 diabetes. A review. *Nutrition, Metabolism, and Cardiovascular Diseases*, *20*(5), 366–375.

Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier, from https://doi.org/10.48550/arXiv.1602.04938.

Shapley, L. S. (1951). Notes on the n-Person Game — II: The Value of an n-Person Game, from https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.

附錄

HighBP（高血壓）	是否有高血壓，0 代表沒有，1 代表有。
HighChol（高血脂）	是否有高血脂，0 代表沒有，1 代表有。
CholCheck（血脂檢查）	5 年內是否有檢查過血脂，0 代表沒有，1 代表有。
BMI（身體質量指數）	身體質量指數（BMI）數值為多少。
Smoker（抽菸）	到目前為止是否抽過至少 100 根或 5 包菸，0 代表沒有，1 代表有。
Stroke（中風）	是否曾中風，0 代表沒有，1 代表有。
HeartDiseaseorAttack（心臟疾病）	是否曾有冠狀動脈心臟病或心肌梗塞，0 代表沒有，1 代表有。
PhysActivity（運動）	除工作之外是否有身體活動或運動，0 代表沒有，1 代表有。
Fruits（水果）	每天是否有吃一份水果或更多，0 代表沒有，1 代表有。
Veggies（蔬菜）	每天是否有吃一份蔬菜或更多，0 代表沒有，1 代表有。
HvyAlcoholConsump（重度飲酒）	每周男性是否飲用超過 14 杯酒、女性是否飲用超過 7 杯酒，0 代表沒有，1 代表有。
AnyHealthcare（醫療保險）	是否有任何健康相關保險，0 代表沒有，1 代表有。
NoDocbcCost（沒錢無法就醫）	是否有因為沒錢無法看醫生，0 代表沒有，1 代表有。
GenHlth（一般身體健康）	自己評分自己大致上的健康，1 代表最好，5 代表最差。
MentHlth（心理健康）	1 個月內有幾天自己認為有壓力、憂鬱或情緒問題。
PhysHlth（生理健康）	1 個月內有幾天自己是受傷或不舒服。
DiffWalk（困難走路）	自己是否認為有走路或爬樓梯困難，1 代表最好，5 代表最差。
Sex（性別）	0 代表女性，1 代表男性。
Age（年齡）	年齡，1:18-24 歲、2:25-29 歲、3:30-34 歲、4:35-39 歲、5:40-44 歲、6:45-49 歲、7:50-54 歲、8:55-59 歲、9:60-64 歲、10:65-69 歲、11:70-74 歲、12:75-79 歲、13:超過 80 歲。
Education（學歷）	最高學歷，1:沒上過幼稚園或任何學校、2:小學、3:國中、4:高中、5:大學 1-3 年級、6:大學 4 年級或畢業
Income（收入）	年收入平均，1:少於 10,000 美金、2:10,000-15,000 美金、3:15,000-20,000 美金、4:20,000-25,000 美金、5:25,000-35,000 美金、6:35,000-50,000 美金、7:50,000-75,000 美金、8:75,000 美金或更多。

本文之程式 QRCORD: https://github.com/Andy051089/UniversityProject