**國立中央大學資訊電機學院資訊工程系**

**資料科學導論期末報告**

Department of Computer Science & Information Engineering

College of Electrical Engineering & Computer Science

National Central University

[Introduction to Data Science](https://ncueeclass.ncu.edu.tw/course/21795) Final Report

**基於機器學習的中風預測系統**

Machine learning-based stroke prediction system

**劉韶颺 潘兆新 劉佳璇**

LIU, SHAO-YANG PAN, ZHAO-XIN LIU, JIA-XUAN

**教授：陳弘軒**

[Hung-Hsuan Chen](https://scholars.ncu.edu.tw/zh/persons/hung-hsuan-chen) Ph.D.

中華民國113年1月 JAN, 2024

**目次**

摘要 .............................................................................................................2

1. 簡介 .........................................................................................................3

2. 資料集分析和處理..................................................................................4

2.1 資料視覺化..................................................................................4

2.2 資料處理 ...................................................................................5

3. 模型訓練 .................................................................................................7

3.1 度量指標 .....................................................................................7

3.2 SMOTE ........................................................................................7

3.3 xgboost .........................................................................................7

3.4 EasyEnsemble ..............................................................................8

4. 訓練結果 ................................................................................................ 9

5. 網站部署 ...............................................................................................14

6. 結論........................................................................................................15

6.1 本研究的貢獻............................................................................15

6.2 可改進方向................................................................................15

參考文獻 ...................................................................................................16

**摘要**

本研究旨在設計一個基於機器學習的中風預測系統，根據使用者輸入的相關資訊，預測中風的可能性。本研究採用了二元分類的方式，使用了Kaggle上的中風預測資料集，對資料進行了初步分析和前處理。接著，爲了克服數據集資料的不平衡，本研究嘗試了不同的方法進行訓練，包括SMOTE、對於不同的類別指定不同的訓練權重、EasyEnsemble，並使用了ROC曲線和AUC值來評估模型的效能。本研究將模型部署在一個網站上，讓使用者可以透過網頁表單輸入自己的資訊，並得到中風的預測結果和相關的建議。代碼可在[**https://github.com/chris-pan-0220/stroke-prediction**](https://github.com/chris-pan-0220/stroke-prediction)上獲得。

**1.簡介**

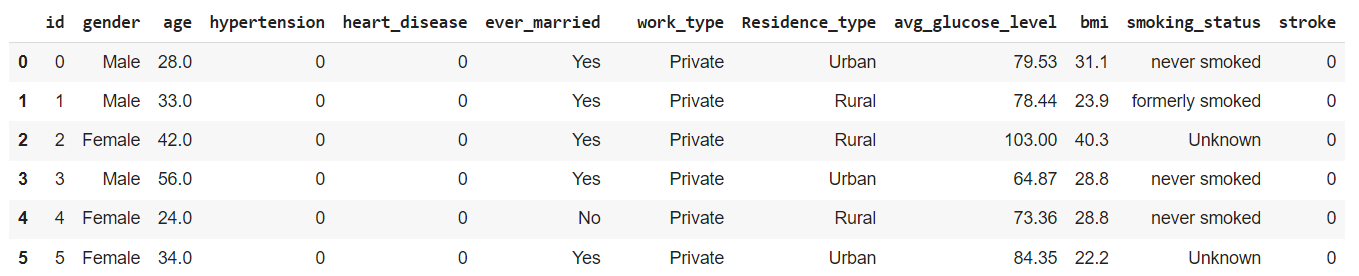
中風是一種常見的腦血管疾病，它發生在腦部的血管被阻塞或破裂，導致腦部缺血或出血，進而造成神經功能障礙或死亡。根據世界衛生組織的統計，中風是全球第二大死因，也是導致殘疾的主要原因之一。因此，預防和治療中風是一個重要的公共衛生課題。然而，並非所有的中風都可以通過預防措施來避免，有些中風是突發的，沒有明顯的先兆或症狀。因此，及早發現中風的跡象，及時就醫，是一種有效的二級預防策略。

隨著人工智能的發展，機器學習作為AI的一個重要分支，已經在各個領域顯示出了強大的應用潛力，機器學習可以處理大量的數據，發現數據中的隱含的模式和關聯，並提供客觀和準確的分析和預測。因此機器學習也可以用於中風的預測，利用患者的基本資訊和健康狀況，計算出中風的可能性，並提供相應的建議。這樣可以幫助高危人群及時發現中風的風險，並採取適當的預防措施，從而降低中風的發生率和死亡率。

本研究的目的是設計和實作一個中風預測系統，利用機器學習的方法，根據使用者輸入的相關資訊，如年齡、性別、職業等，預測中風的可能性，並提供一個簡單的網站介面，讓使用者可以方便地使用本研究的成果。

**2.資料集分析和處理**

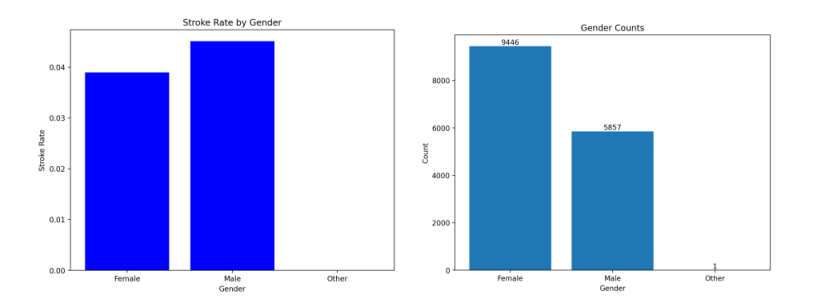
本研究使用了Kaggle上的一個公開資料集，包含了15,304筆病患的相關資訊，包含12個特徵欄位。

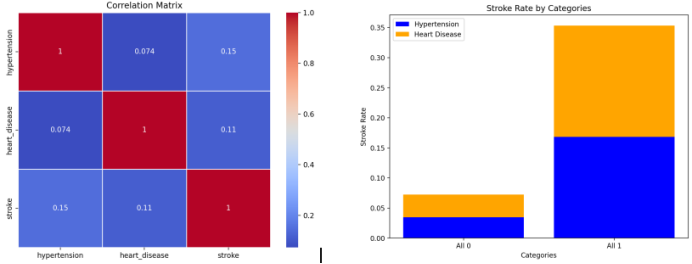


**圖 1 資料集信息**

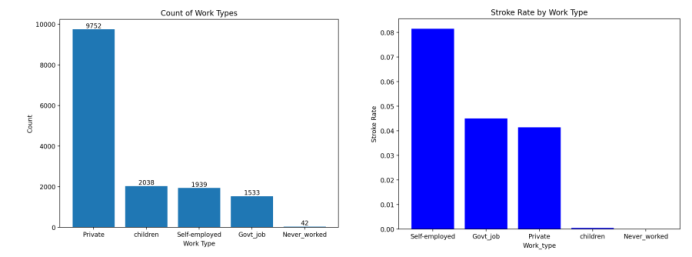
本研究對資料進行了以下幾個步驟的分析和處理：

**2.1資料視覺化**

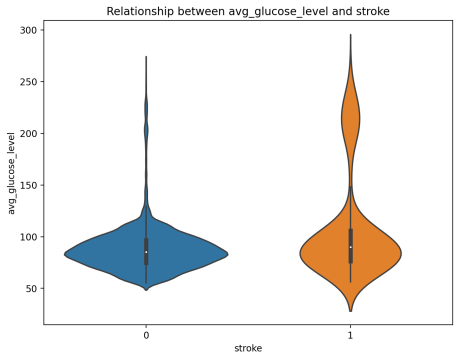
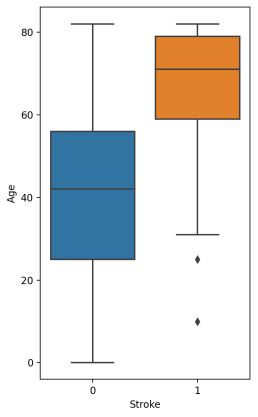
**圖2 性別與中風的關係**

****

**圖3高血壓和心臟病與中風的關係**



**圖4 工作類型與中風的關係**



**圖5 年齡與中風的關係 圖6 平均血糖值與中風的關係**

由圖2可以看出資料集中風的群體中，男性比女性略高，且性別為other的群體只有一筆資料，因此移除other；由圖3可以看出分別患高血壓和分別患心臟病的人中風的相關係數很小，但是同時換高血壓和心臟病的人中風的概率明顯升高；由圖4可以看出雖然大多數人的工作型態都是private，但是Self-employed中風的比例稍微高一些；由圖5可以看出年齡越高，中風的概率越大；由圖6可以看出沒中風的人主要血糖值主要都分布在平均值上；而中風人的主要血糖值除了分布在平均值上，也分布在血糖值偏高的部分。

**2.2資料處理**

* **選取代表性對象**：

16歲以下的樣本比較沒有代表性，因爲一般認為心臟疾病、高血壓與中風會比較有相關性，但是對於幼齡兒童來說，我們不知道做這些檢測的可信度是否足夠，而對於稍長一點的青少年，此兩種疾病是完全沒有，將資料納入訓練可能會影響結果。所以本研究選取了16嵗以上的人作爲樣本數據。

* **極端值處理**：

本研究發現資料集中有一些極端值，如BMI達到80以上，這些值可能是輸入錯誤或者是異常個案，會影響模型的準確性，因此本研究將binning來減少極端值的影響，只保留正常範圍內的資料。

* **編碼**：

本研究對一些類別型的特徵進行了編碼，如性別、工作型態、居住地和吸菸狀況，將它們轉換成數值型的特徵，以便模型可以處理。本研究使用了one-hot encoding的方法，將每個類別轉換成一個二元的特徵，表示該類別是否存在。

**3.模型訓練**

爲了克服數據集資料的不平衡，本研究嘗試了不同的方法進行訓練，包括SMOTE、對於不同的類別指定不同的訓練權重、EasyEnsemble。並使用了AUC作為評估模型性能的指標。

**3.1 度量指標**

**ROC(Receiver operator characteristic）曲線**：ROC空間將偽陽性率（FPR）定義為 X 軸，真陽性率（TPR）定義為 Y 軸。TPR表示在所有實際為陽性的樣本中，被正確地判斷為陽性之比率。FPR表示在所有實際為陰性的樣本中，被錯誤地判斷為陽性之比率。ROC曲線反映了不同閾值下，分類器的敏感度和特異度的變化情況。

**AUC(Area under the Curve of ROC)**：ROC曲線下方的面積，其意義為：

AUC = 1，是完美分類器，採用這個預測模型時，存在至少一個閾值能得出完美預測。

0.5 < AUC < 1，優於隨機猜測。這個分類器（模型）妥善設定閾值的話，能有預測價值。

AUC = 0.5，跟隨機猜測一樣，模型沒有預測價值。

AUC < 0.5，比隨機猜測還差；但只要總是反預測而行，就優於隨機猜測。

**3.2 SMOTE**

SMOTE是一種對少數類別進行上採樣的方法，它可以通過在少數類別的樣本之間插值生成新的合成樣本，從而增加少數類別的數量，平衡類別分佈。該方法的步驟如下：

1. 設定一個採樣倍率 N，也就是對每個樣本需要生成幾個合成樣本

2. 設定一個近鄰值 K ，針對該樣本找出 K 個最近鄰樣本並從中隨機選一個

3. 根據公式來創造 N 個樣本

**3.3 Xgboost**

xgboost是一種基於梯度提升決策樹的集成學習方法，它可以自動處理缺失值和類別變量，並且具有高效、靈活和可擴展的特點。我們使用了xgboost內置的scale\_pos\_weight參數來處理類別不平衡的問題，該參數可以根據類別的比例調整損失函數的權重。

**3.4 EasyEnsemble**

EasyEnsemble是一種對多數類別進行下採樣的方法，先將資料切分成好數個子集，並且對多數標籤做隨機採樣，採樣次數等於子集數量，並且採樣數量等於少數類別的數量。將採樣後的多數類別資料，以及少數類別資料放入每一個子集中，對於每一個子集進行使用不同的分類器進行ensemble learning。進行預測時，會將資料輸入每一個分類器，並將這些預測結果進行加權。

**4.訓練結果**

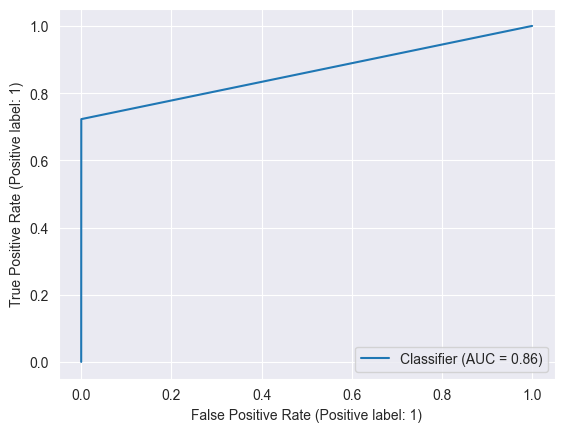
本研究採用的baseline是Xgboost，由圖7可知在測試集上的ROC curve幾乎趨近於random classifier，無法很有效的處理不平衡的數據集。

Model-l SMOTE+Xgboost是對訓練集採用SMOTE進行上採樣，然後再使用Xgboost進行訓練。如圖9、10雖然在訓練集上的ROC curve有顯著的提升，但是在測試集上的ROC curve一樣幾乎趨近於random classifier，代表加入SMOTE方法依然無法很有效的處理不平衡的數據集。

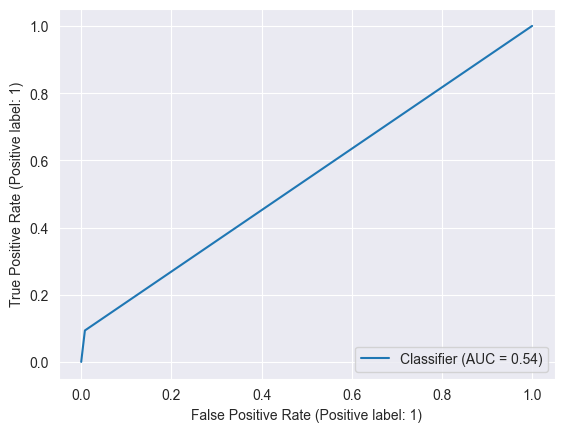
Mode-2 Xgboost+scale\_pos\_weight是使用Xgboost進行訓練，並且使用scale\_pos\_weight參數來調整訓練時兩種類別的權重比例。從下圖11、12可以看到在測試集上的ROC curve表現有顯著的提升，此一方法能夠讓模型成功學習不平衡的數據。

Model-3 EasyEnsemble使用多個分類器對不平衡數據集進行採樣並進行集成學習，由圖13、14的訓練集、測試集結果來看，都有最佳的表現。

**Baseline: Xgboost**

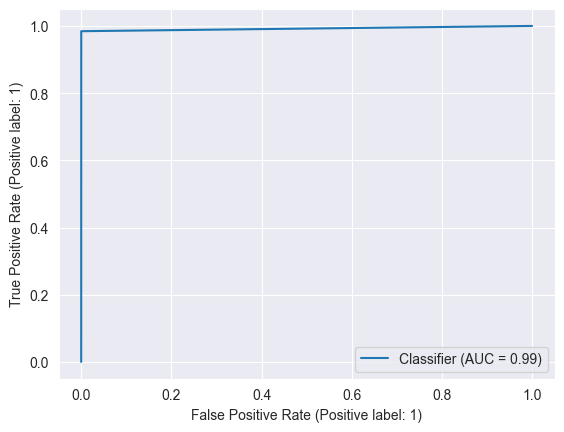
****

**圖7 train-baseline**

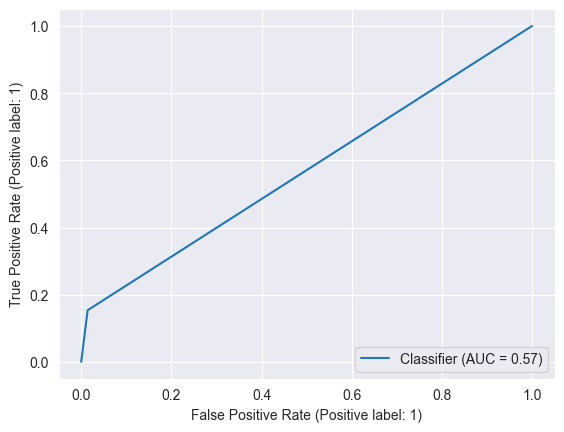


**圖8 test-baseline**

**Model-1: SMOTE + Xgboost**

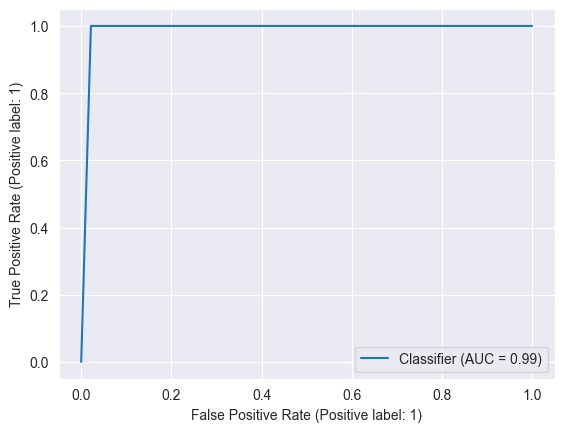


**圖9 train-Xgboost+SMOTE**

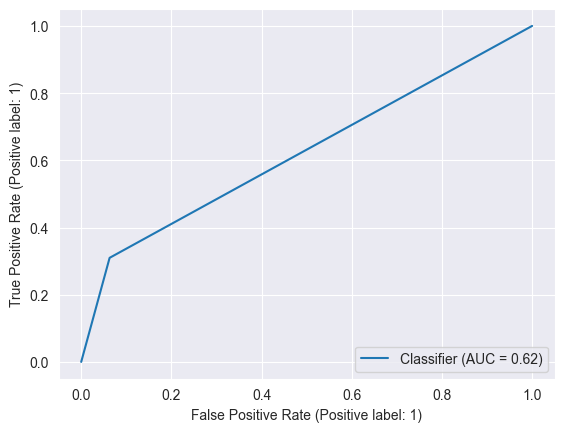


**圖10 test-Xgboost+SMOTE**

**Model-2: Xgboost + scale\_pos\_weight**

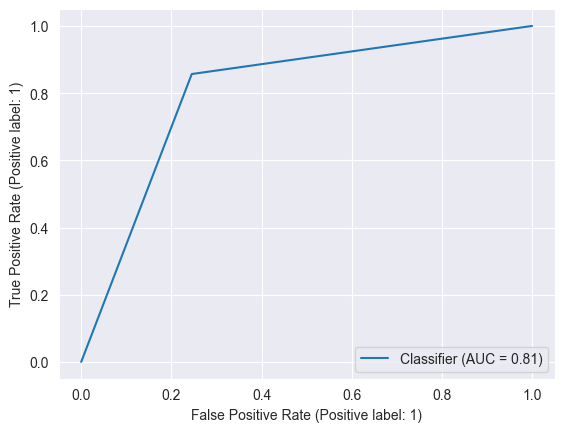


**圖11 train-Xgboost+scale\_pos\_weight**

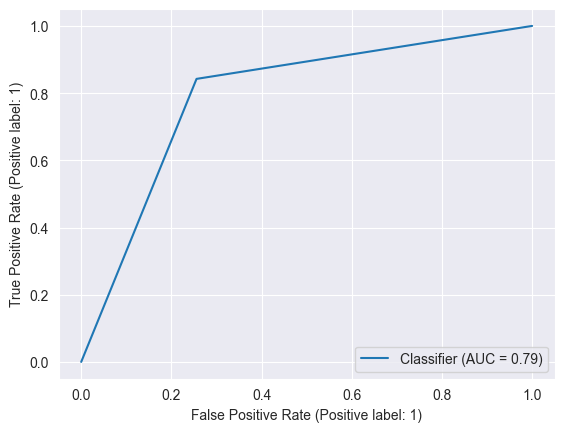


**圖12 test-Xgboost+scale\_pos\_weight**

**Model-3: EasyEnsemble**



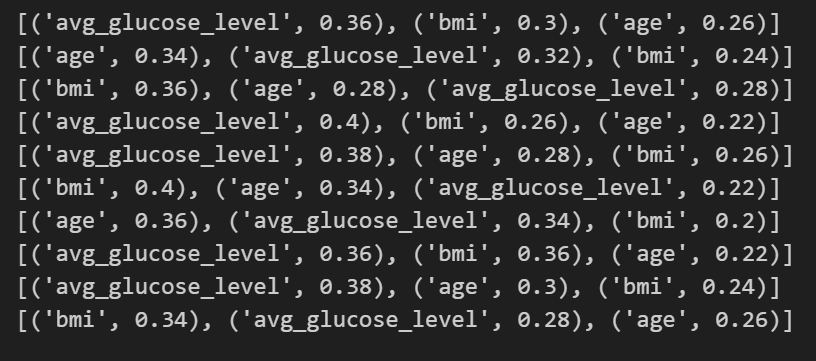
**圖13 train-EasyEnsemble**



**圖14 test-EasyEnsemble**

本研究將三種模型的預測結果進行了比較和分析，發現EasyEnsemble的模型表現最好。本研究認為，這是因為EasyEnsemble的模型可以有效地解決資料的不平衡問題，並且可以利用多個分類器的集成學習來提高預測的準確性和穩定性。

**feature importance**

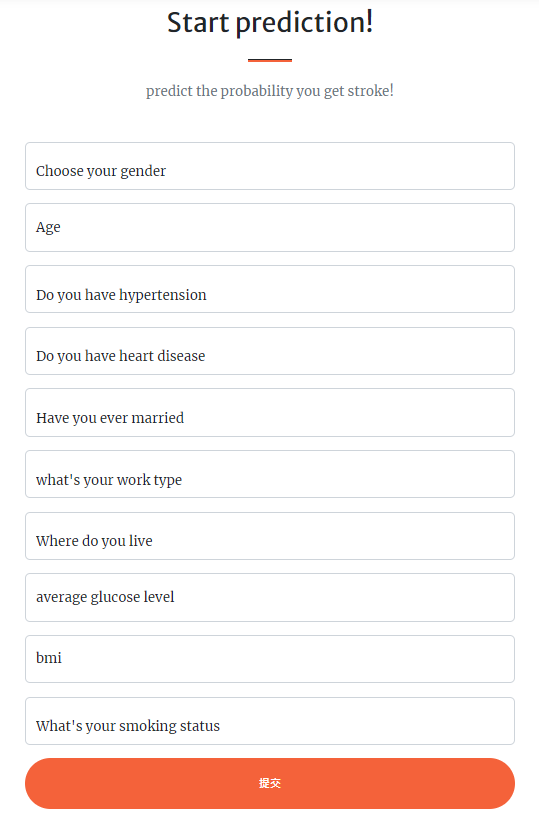


**圖15 feature importance**

由圖15，每一個classifier的feature importance都不太相同, 但是重要的feature大致一樣.縂體來看，有三個特徵起到決定性的作用，分別是：**age**、**avg\_glucose\_level**、**bmi**。

**5.網站部署**

為了讓使用者可以方便地使用我們的模型，我們將模型部署在一個網站上，讓使用者可以透過網頁表單輸入自己的相關資訊，並得到中風的預測結果。我們使用了Flask作為網站框架，並使用了Bootstrap作為網頁設計。



**圖16.網站使用界面截圖**

使用者可以在表單中填寫自己的年齡、性別、職業、血壓、心臟病、血糖、BMI、吸菸狀況等資訊，並點擊提交按鈕，就可以得到我們的模型的預測結果和相關的建議。模型會根據使用者的資訊，計算出中風的可能性，並將其顯示在網頁上。

## **6.結論**

在本研究中，我們設計和實作了一個基於機器學習的中風預測系統，利用EasyEnsemble和XGBoost的方法，根據使用者輸入的相關資訊，預測中風的可能性。我們使用了一個Kaggle公開的中風預測資料集，對資料進行了分析和前處理，並使用了AUC作為模型的評估指標。我們的實驗結果顯示，我們的模型在處理資料類別不平衡的問題上有顯著的改善，並且對正負類別有良好的區分能力。我們還對我們的模型進行了特徵重要性的分析，發現年齡、平均血糖值和BMI是最重要的三個特徵，與常識和文獻中的發現相一致。本研究的模型部署在一個網站上，讓使用者可以方便地使用模型，並得到中風的預測結果和相關的建議。

**6.1本研究的貢獻：**

使用了EasyEnsemble的方法，有效地處理了資料類別不平衡的問題，並提高了模型的預測準確性和穩定性。

對模型進行了特徵重要性的分析，發現了影響中風的風險的主要因素。

將模型部署在一個網站上，讓使用者可以方便地使用模型，並得到中風的預測結果。

**6.2可以改進的方向：**

可以嘗試使用不同的資料集來訓練模型，並且使用更多的特徵來提高模型的預測能力。也可以嘗試使用更多的方法來處理資料類別不平衡的問題，如權重調整、成本敏感學習等，並且比較不同方法的效果和優缺點。

**參考文獻**

<http://ntur.lib.ntu.edu.tw/bitstream/246246/160578/1/56.pdf>

<https://www.commonhealth.com.tw/book/66>

<https://www.cmuh.cmu.edu.tw/HealthEdus/Detail?no=4771>

<https://www1.cgmh.org.tw/strokelnk/07/20160603/20160603_1.pdf>

<http://www.dmcare.org.tw/up3/2005-4-%E6%88%B4%E9%81%93%E6%81%A9.pdf>

<https://getbootstrap.com/docs/5.3/getting-started/introduction/>