DIABETES PREDICTION

Using Machine Learning Model

Group 1

110072250陳冠倫

109034057馮棨晟

109011261謝裕紀

TABLE OF CONTENTS

Motivation

為什麼糖尿病要進行預測？

- 糖尿病是一種越早發現徵兆就越容易控制的疾病。若能在早期就進行飲食管理或接受治療，將能大幅降低併發症的風險，進而有效減少死亡率與長期醫療成本。
- 糖尿病因為需要抽血檢查費用不低。對於沒有健保的民眾，或對整體健保體系而言，全面性檢測所帶來的財務負擔都不小。
- 因此，若能透過較為簡單、低成本的健康指標來預測個體罹患糖尿病的風險，不僅能協助更有效率地篩檢潛在患者，也有助降低健保成本壓力，進一步提升整體公共衛生的成效。

Data Exploration

DATA SOURCE

- 本資料集來自美國疾病控制與預防中心（CDC）於2015年進行的「行為風險因素監測系統」（BRFSS）調查，該調查是全球最大的健康相關電話調查系統，旨在收集美國成人的健康風險行為、慢性健康狀況和使用預防服務的情況。
- 資料集包含 253,680 筆樣本，共有 22 個欄位，其中包含 21 個作為模型輸入的特徵變數，以及 1 個目標變數 Diabetes_binary，代表個體是否罹患糖尿病。

Binary Columns - 1

Diabetes_binary	是否有糖尿病(target)
HighBP	是否有高血壓
HighChol	是否有高膽固醇
CholCheck	是否有在5年內檢查過膽固醇指數
HvyAlcoholConsump	是否為重度酒精
Smoker	人生中是否抽過超過100根菸
Stroke	是否曾經中風
HeartDiseaseorAttack	是否曾有心肌梗塞相關疾病

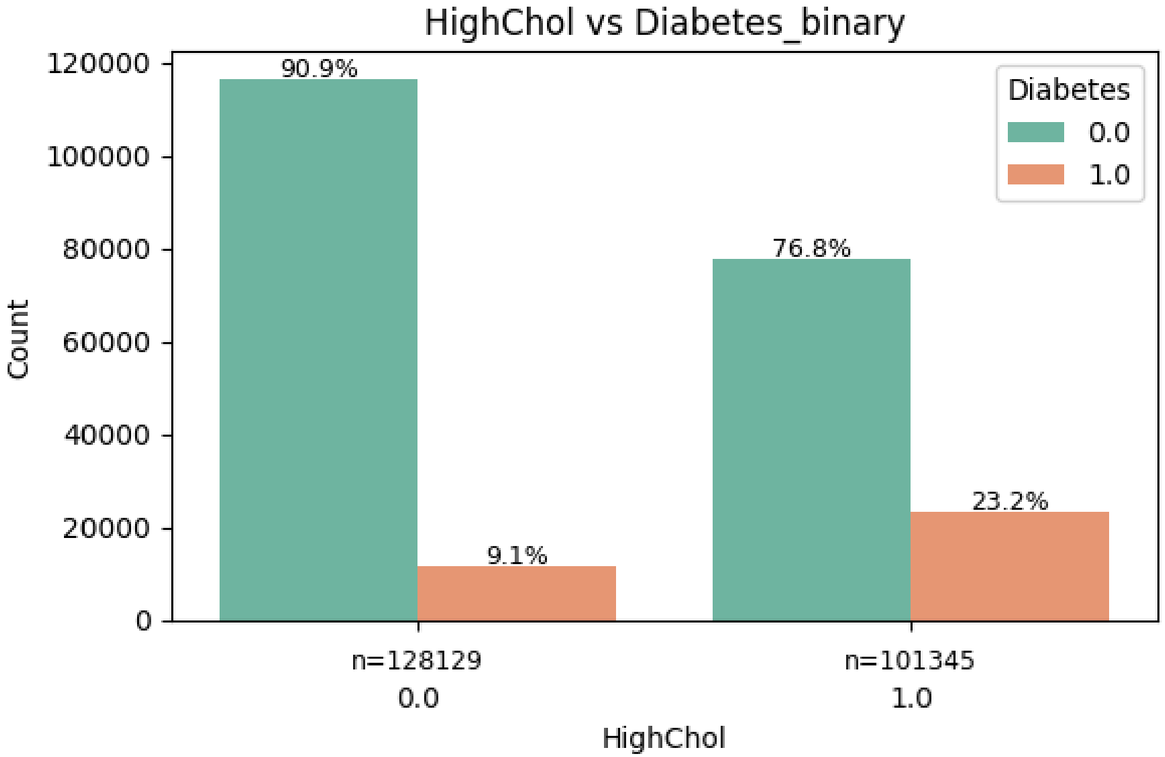
Binary Columns - 2

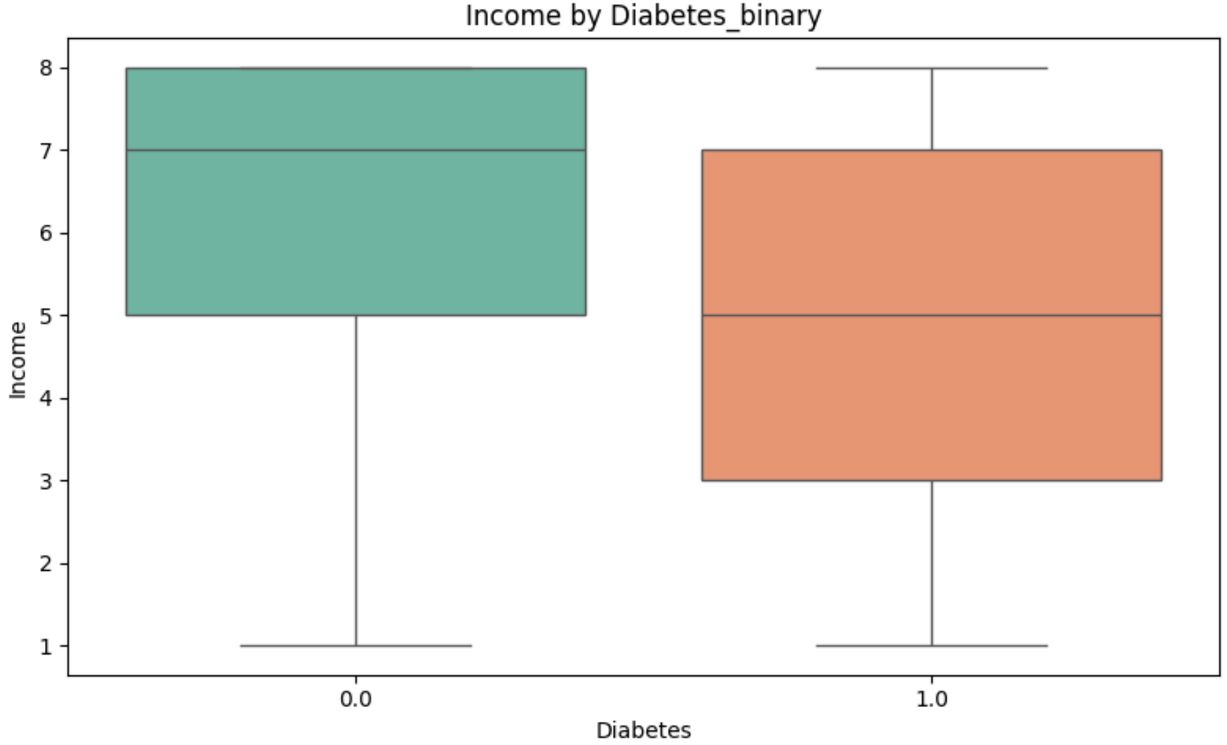
PhysActivity	過去30天是否有運動
Fruits	是否每日攝取蔬菜
Veggies	是否每日攝取蔬菜
AnyHealthcare	是否擁有任何健康保險
NoDocbcCost	過去一年是否曾因費用而無法就醫
DiffWalk	走路與爬樓梯是否感到困難
Sex	性別

Numeric Columns

BMI	身體質量指數
GenHlth	自評健康狀況 (1-5分)
MentHlth	過去30天有幾天感到壓力與情緒問題
PhysHlth	過去30天有幾天身體健康有問題
Age	年齡 (從18到80歲以上 分為13個區間)
Education	教育程度 (從未受教育到大學以上，分為6個區間)
Income	個人年收入 (分為8個區間)

Bar Chart / Boxplot





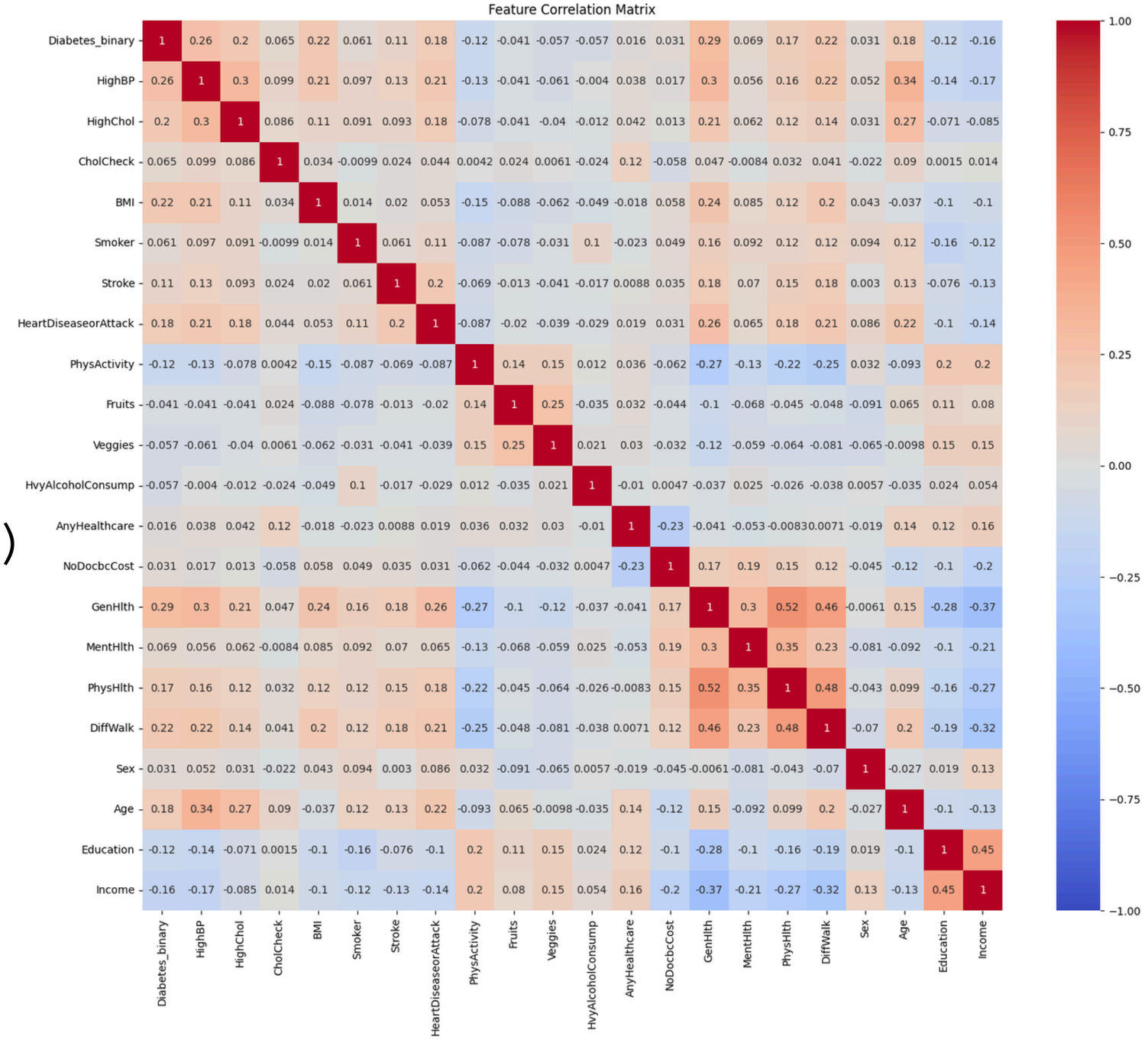
Heat Map

變數之間的相關度都不高

Feature之間最大的只到0.52

(自評心理狀況/自評身體狀況)

共線性問題不大

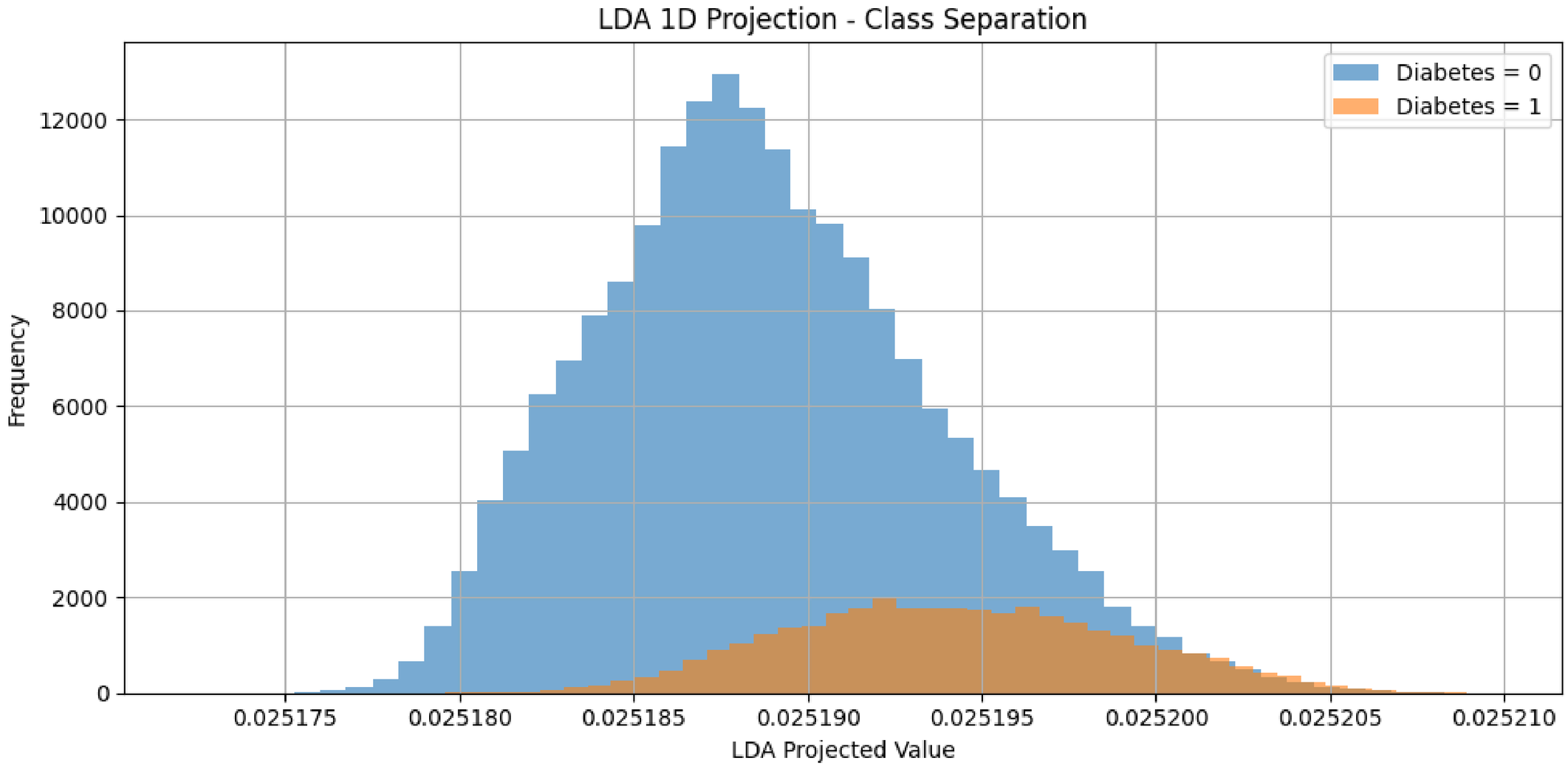


Data Preprocessing

- 資料沒有缺失值要補
- 將binary和numeric欄位分開處理
  - 將numeric欄位做正規化
  - binary則保持原狀
- Train: Validation: Test = 6:2:2

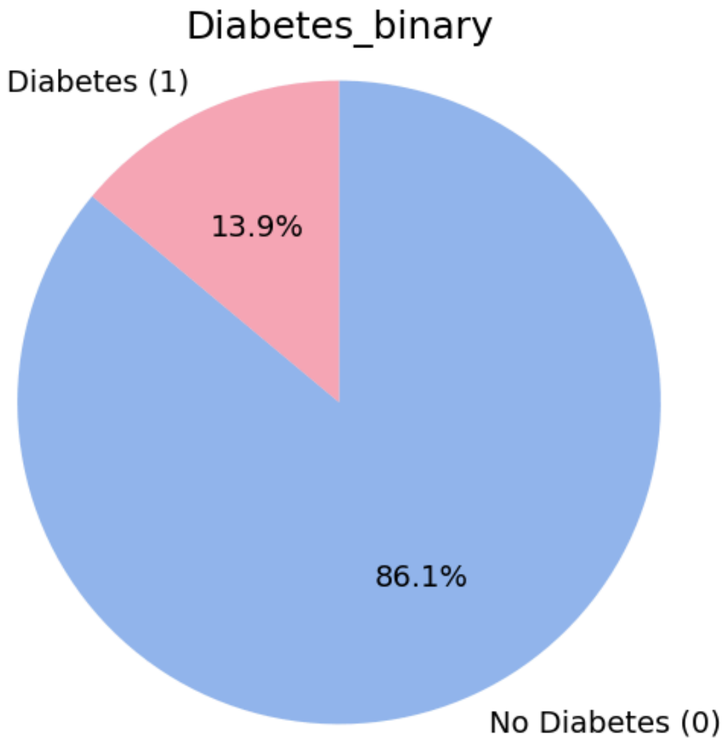
Dimension Reduction

LDA



Modeling

Model Evaluation & Threshold Tuning



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

由於目標 Diabetes_binary 非常不平衡，因此Accuracy 無法反映模型的好壞程度，需要其他指標。

		Predicted 0	Predicted 1
Actual	0	TN	FP
	1	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

- Recall:
  - 有糖尿病的人是否都有被模型準確找到
- Precision:
  - 模型判斷為糖尿病的人裡有多少是真的

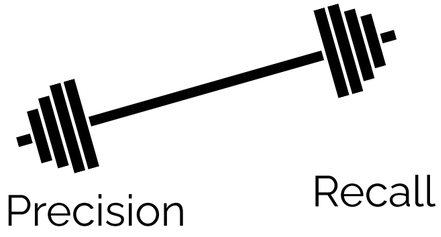
Threshold Tuning

Fβ Score

$$F_1 = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

$$F_\beta = \frac{1 + \beta^2}{\frac{\beta^2}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

F0.5

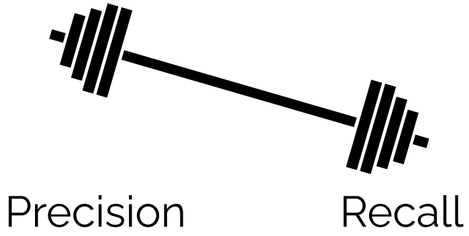


- Emphasis on reducing FP

F1

- Goal: balanced performance of FP & FN

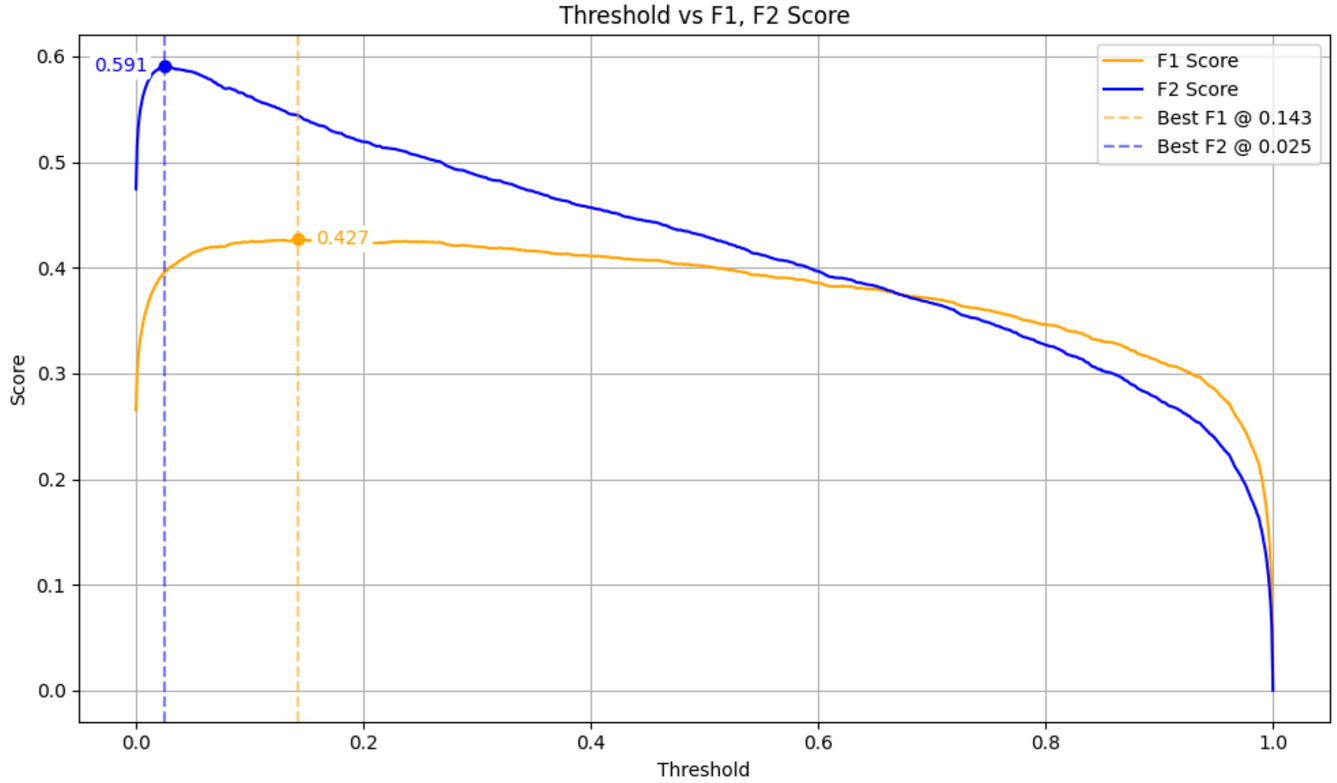
F2



- Emphasis on reducing FN

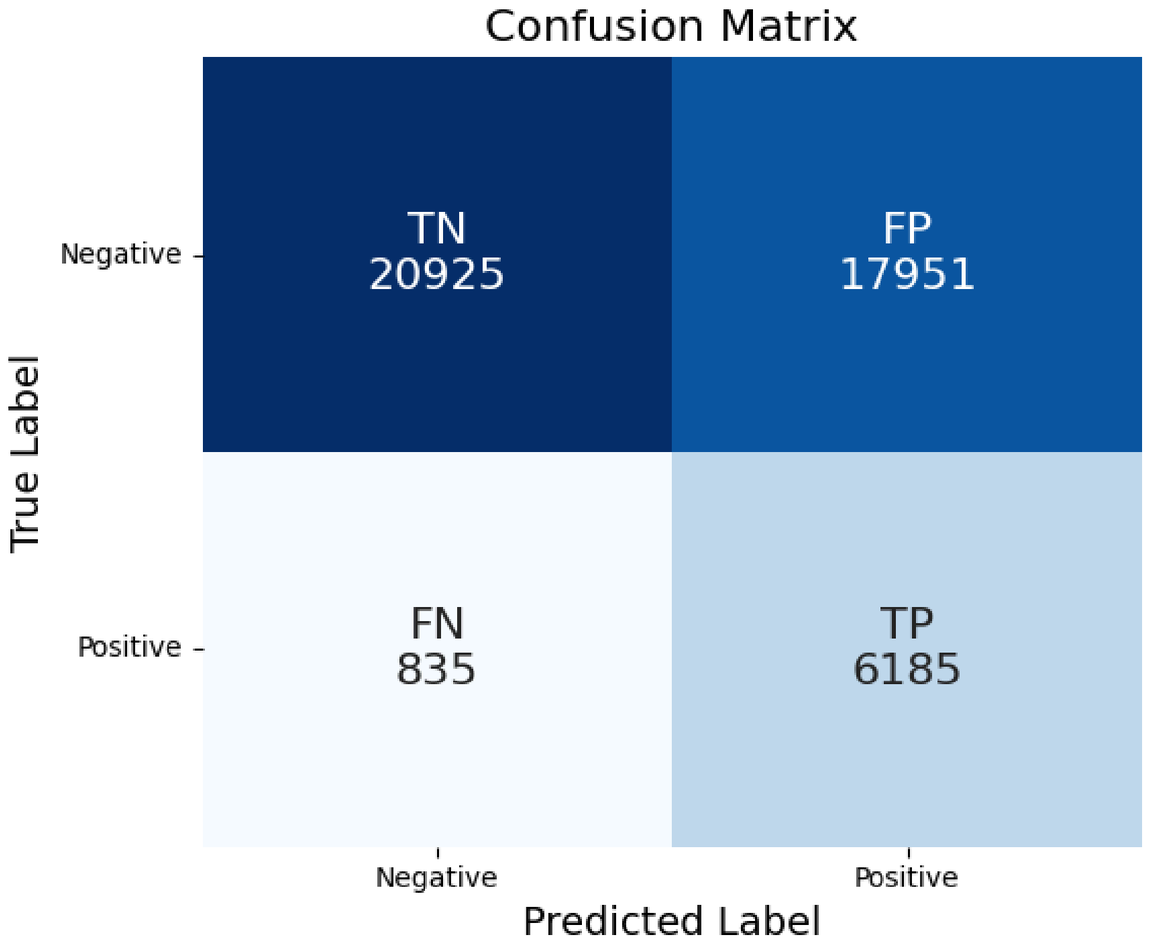
Naive Bayes

- Mixed naive bayes: Gaussion + Bernoulli
- Train: Val: Test = 6:2:2
- Max F2 (**validation**): 0.591
- Threshold: 0.025



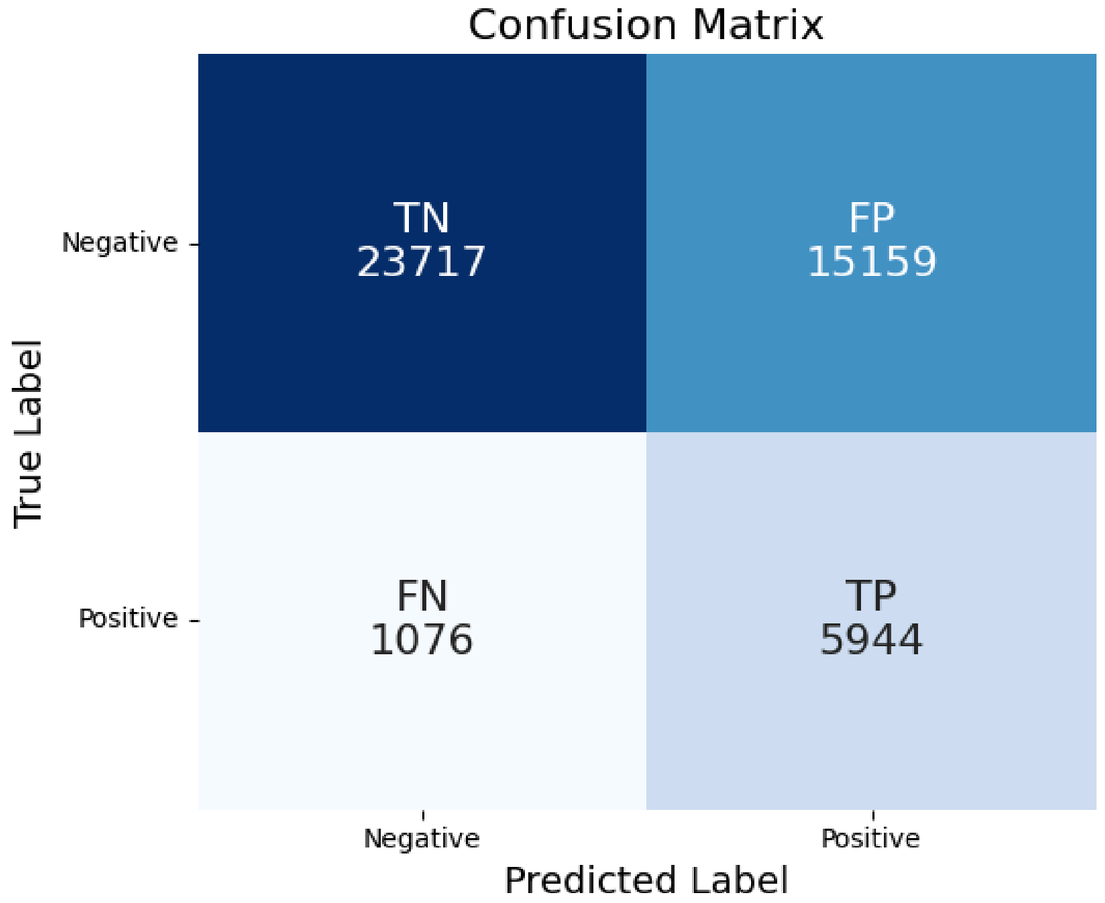
Naive Bayes

- **Threshold: 0.025**
- Accuracy: 0.589
- Precision: 0.256
- Recall: 0.881
- F1: 0.397
- **F2: 0.593**



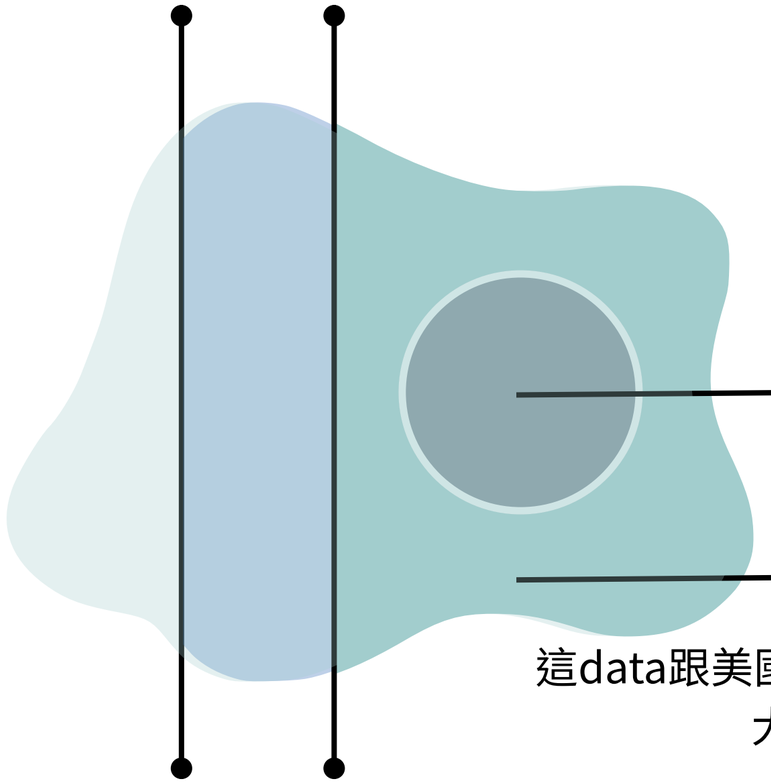
Least Square

- **Threshold：0.161**
- Accuracy: 0.646
- Precision: 0.282
- Recall: 0.847
- F1: 0.423
- **F2: 0.604**



Gradient Descent

評估　驗證　訓練



陽：陰 = 1：1

陽：陰 = 84：16

這data跟美國糖尿病患者比例差不多
大概14%左右

Test : Validation : Training = 2 : 2 : 6

模型：Logistic Regression（no class weight)
損失函數：Log Loss, Binary Cross Entropy
epoch(學習次數)= 10000

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

y_i：真實標籤（0 或 1）

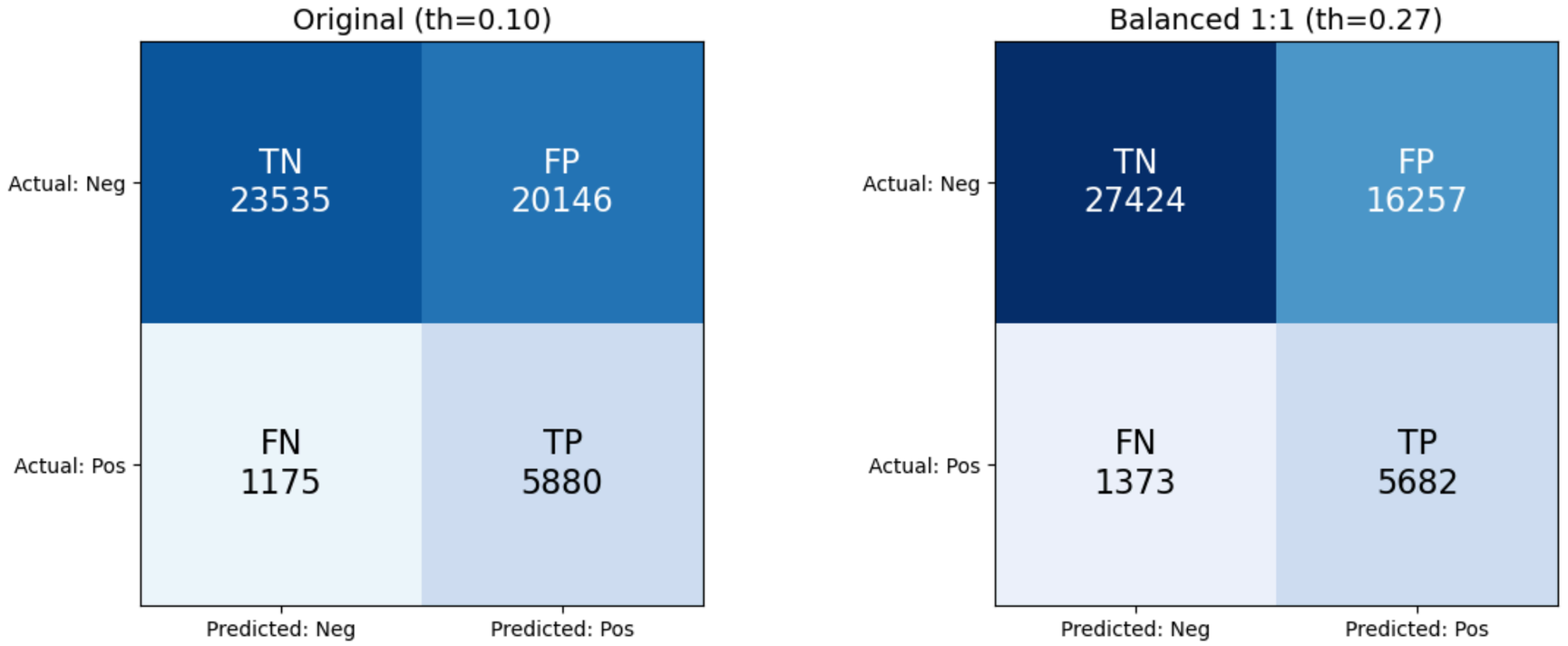
p_i：模型輸出的機率（$\sigma(X_i w)$，即 sigmoid 函數的輸出）

N：樣本數

目的：為了screening(更準確的篩選陽性)
提升“F2 score”

方法：**1. Data Sampling (only test data)**
2. 找到適合的閾值

問題：陰性陽性比例不均勻，導致模型“亂猜”陰性 → **Random Undersampling**



Original Train Data (best th)
Threshold: 0.10

Accuracy: 0.5798
Precision: 0.2259
Recall: 0.8335
F1: 0.3555
F2: 0.5420

Balanced Train Data (best th)
Threshold: 0.27

Accuracy: 0.6525
Precision: 0.2590
Recall: 0.8054
F1: 0.3919
F2: 0.5664

Our Findings

- Naive Bayes, Least Square, Gradient Descent 等模型在調整閾值後皆可有效提升 Recall（從 0.81 提升至 0.83），成功降低偽陰性。然而，Precision 會同時下降至約 0.28，偽陽性略有增加。
- 為了平衡兩者，我們採用**F2分數**進行模型評估。
- 在本資料與模型結構下，F2 分數的上限約為 0.59，若欲進一步提升需仰賴更複雜模型、更多特徵，或其他演算法。再者，這份資料來自電話問卷，特徵受限於調查內容而非醫療客觀判斷，因此存在預測力上的天花板。
- 總結來說，對於我們的目標是避免漏判、提高 Recall，這些方法皆具實用性。

成效

Naive Bayes F2=0.593
Least Square F2=0.603
Gradient Descent F2=0.566← ?
各種方法的 F2 分數（重視 Recall）都差不多，約為 0.59

進步空間

Gradient Descent (分類)

- **換到別的Model**
- **加上class weight**
- **改Sampling的方式**
  Random→ Near Miss

https://www.ibm.com/think/topics/downsampling

Thank You