糖尿病分析

Yi-Hong Wang

2025-09

1 介紹

1.1 前言

糖尿病作為我國 10 大主要死因之一，並且具有 250 萬病友的國主要慢性病之一，如何從身體檢測當中提早預知到何種數據可能與糖尿病相關變得重要。因此我決定嘗試從著手分析 kaggle 上糖尿病病人的資料，並尋找出與糖尿病具有最強關聯的檢測數據。

1.2 資料簡介

資料來自kaggle，為糖尿病人的檢測資訊以及糖尿病患病狀態，共有 264 位病人，128 位患有糖尿病，96 位沒有糖尿病，以及 40 位糖尿病前期，此外還有每位病人有 12 筆檢測數據作為變數，變數說明如下：

1. Gender: 個體的生理性別。通常編碼為：0 = 女性，1 = 男性。性別可能會因荷爾蒙和生活方式的差異影響糖尿病風險。
2. AGE: 個體的年齡。年齡是糖尿病風險的一個關鍵因素，尤其是 45 歲以上，隨著年齡增長風險會增加
3. Urea: 血液中尿素含量的測量 (單位: mg/dL)。高尿素水平可能表明腎臟問題，這是糖尿病常見的併發症。正常範圍： 7–20 mg/dL。
4. Cr: 測量血液中肌酐酸的水平 (單位: mg/dL)。也是腎臟功能的標誌物。肌酐酸水平升高可能表明腎功能受損，這通常與糖尿病有關。正常範圍： 0.6–1.3 mg/dL
5. HbA1c: 一個關鍵指標，代表過去 2-3 個月內的平均血糖水平，以百分比表示。正常：<5.7% - 糖尿病前期：5.7–6.4% - 糖尿病： 6.5%。
6. Chol: 血液中的總膽固醇含量。高膽固醇是心血管疾病的風險因素，通常在糖尿病患者中可見。正常：<200 mg/dL。
7. TG: 測量血液中脂肪的含量 (單位: mg/dL)。高水平與胰島素抵抗和代謝症候群有關。正常：<150 mg/dL。
8. HDL:「好」膽固醇 (單位: mg/dL)。水平越高越好。它有助於清除血液中多餘的膽固醇。理想值：>40 mg/dL (男性)，>50 mg/dL (女性)。
9. LDL:「壞」膽固醇 (單位: mg/dL)。高水平會導致斑塊積聚在動脈中。最佳值：<100 mg/dL。
10. VLDL: 另一種「壞」膽固醇 (單位: mg/dL)。主要攜帶三酸甘油酯。通常根據 TG/5 估算。高 VLDL 與糖尿病風險增加有關。正常：2–30 mg/dL。
11. BMI: 根據身高和體重測量身體脂肪 (單位: kg/m^2)。肥胖 (BMI 30) 是第二型糖尿病的主要風險因素。正常：<18.5 –體重過輕，18.5–24.9 –體重正常，25–29.9 –超重， 30 –肥胖。
12. Class: 目標標籤，表示糖尿病狀態。編碼為：0 = 非糖尿病，1 = 糖尿病，2 = 糖尿病前期。這是您要預測或分類的結果。

1.3 模型簡介

在此使用了 2 種模型，分別是 Random forest 以及 Xgboost 模型

1.3.1 隨機森林（Random Forest)

核心概念：

隨機森林是一種基於「群體智慧」的演算法，它屬於集成學習 (Ensemble Learning) 的範疇。它的基本思想是，由一群相對簡單的個體（決策樹）做出預測，再將所有個體的預測結果整合起來，得到最終的結果。

運作方式：

- 隨機抽樣資料：從原始資料集中隨機抽取多個子樣本集。
- 隨機選擇特徵：在訓練每一棵決策樹時，只會從所有特徵中隨機選擇一部分來進行分裂。
- 多棵決策樹：根據上述兩點，分別建立多棵彼此獨立且隨機的決策樹。
- 結果整合
  - 分類問題：所有決策樹進行「投票」，以多數決的方式決定最終的類別。
  - 回歸問題：所有決策樹的預測結果取「平均」，作為最終的預測值

1.3.2 XGBoost（Extreme Gradient Boosting)

核心概念：

XGBoost 也是一種強大的集成學習演算法，它屬於梯度提升 (Gradient Boosting) 的家族。它的核心思想是「逐步優化」，每一棵新樹的建立都是為了彌補前面所有樹的預測誤差，將模型一步一步推向更好的方向。

運作方式：

- 建立第一棵樹：先建立一棵簡單的決策樹
- 計算殘差：計算第一棵樹的預測值與實際值之間的「殘差」(也就是預測錯誤)。
- 建立新樹：建立第二棵樹來預測並修正這些殘差，然後再計算新的殘差。
- 逐步迭代：不斷重複這個過程，每一棵新樹都專注於修正前一棵樹的錯誤，直到殘差達到一個預設的最小值，或者達到指定的樹木數量。

2 資料分析

2.1 探索資料分析

繪製出 BMI 在不同 Class 下的分配

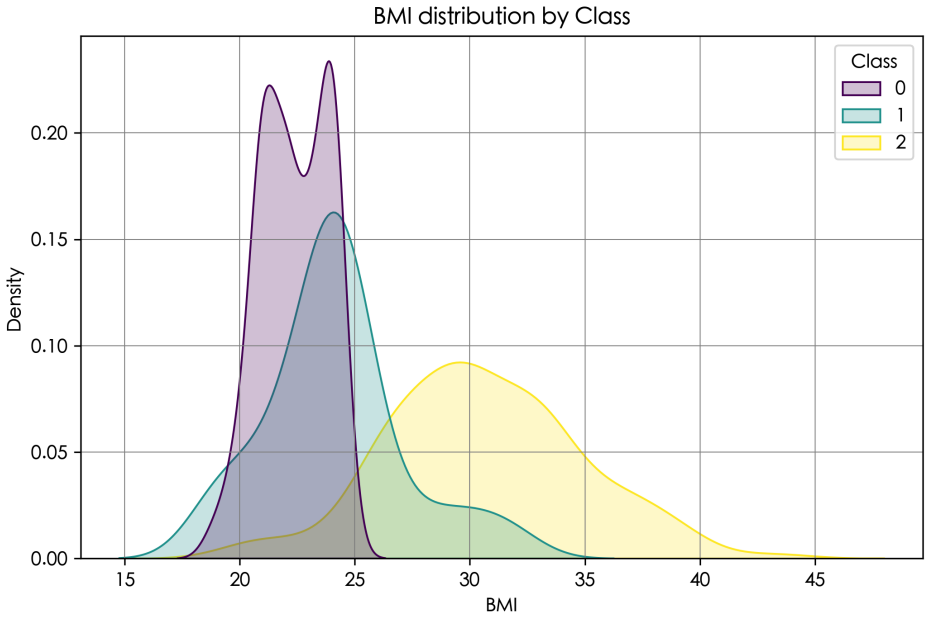

Figure 1

可以看出第 2 組的 BMI 分佈明顯與其他 2 組有明顯差異，糖尿病前期的 BMI 分佈集中在較高的值。接著再繪製出不同 Class 間 HbAlc 的分佈，

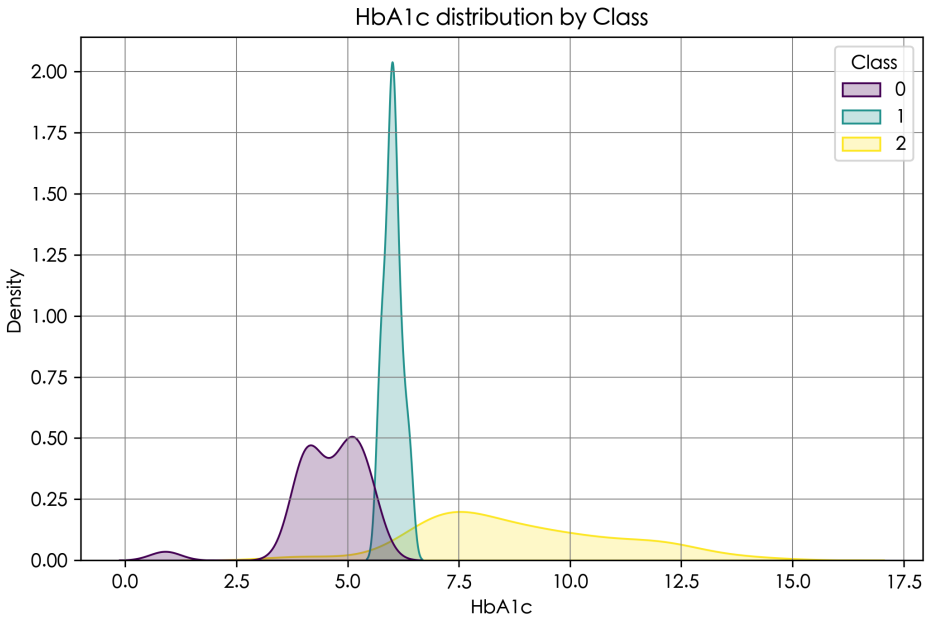

Figure 2

可以看出 3 組的 BMI 分佈都明顯與其他 2 組有明顯差異，對於分類糖尿病很可能來說為重要變數最後畫出 HbAlc 與 BMI 的散佈圖並為不同 Class 加上不同顏色，

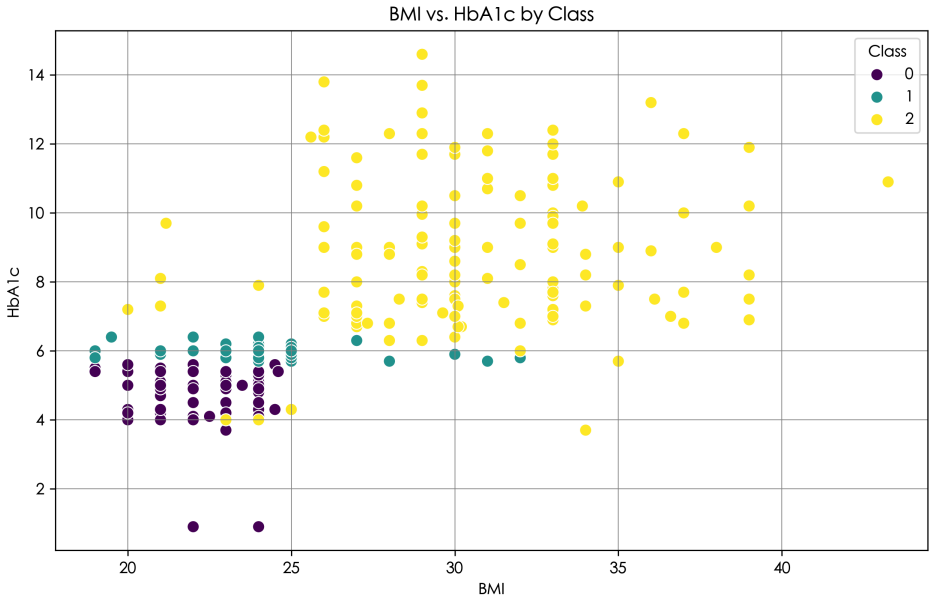

Figure 3

可以看出 Class 可被這 2 個變數分割成功，接著我們建立分類預測模型

2.2 建立模型分析

2.2.1 隨機森林（Random Forest）

模型調控參數後擬合結果模型預測的分類準確率為 96%，並且畫出模型預測結果的 confusion matrix

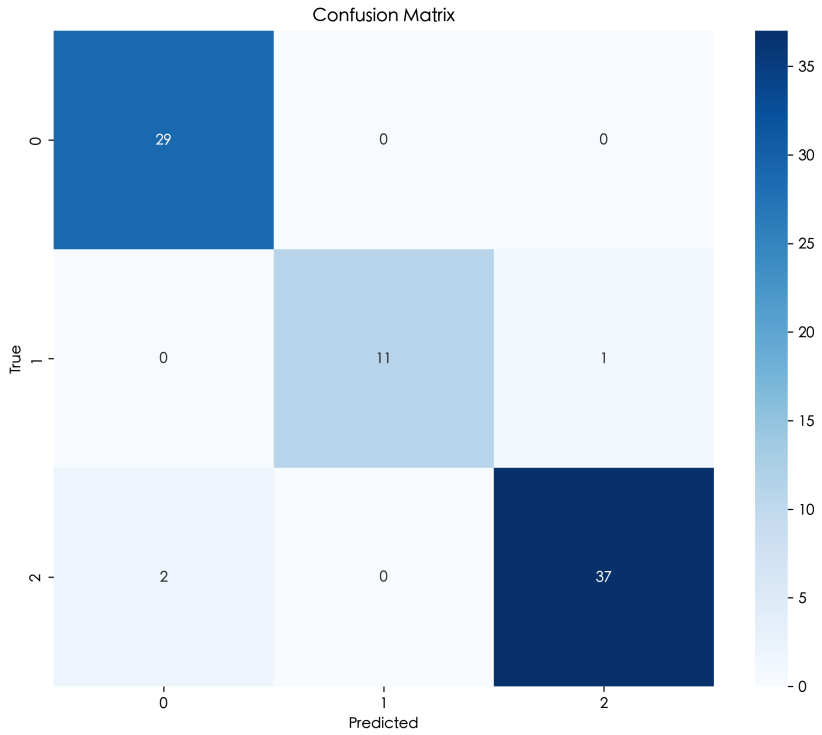

Figure 4

模型表現良好，只有 2 個樣本預測錯誤，對於糖尿病 (Class = 1) 的預測準確度為 100%。接著畫

出 Random forest 模型中的變數重要程度。

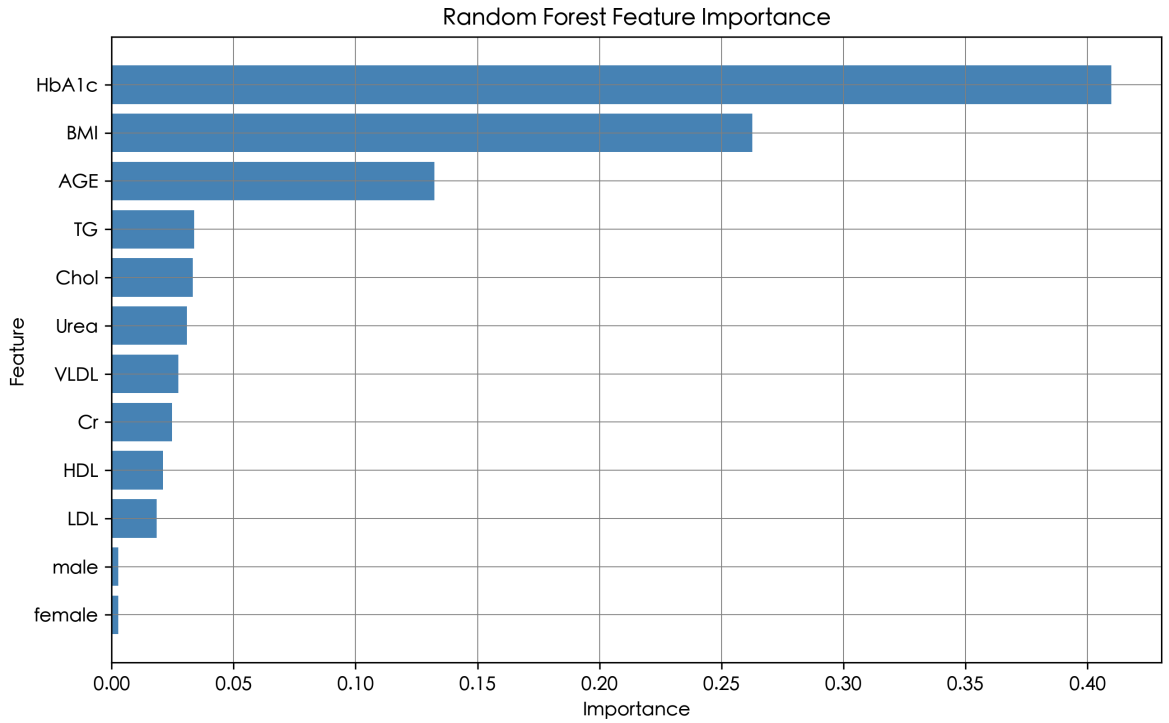

Figure 5

如同前面 Exploratory Data Analysis 的結果，糖尿病的分類預測基本上可由 2 個變數 HbAlc 與 BMI 決定。

為了更詳細的看出 Random forest 如何根據變數分類，畫出模型的 PartialDependenceDisplay 觀察

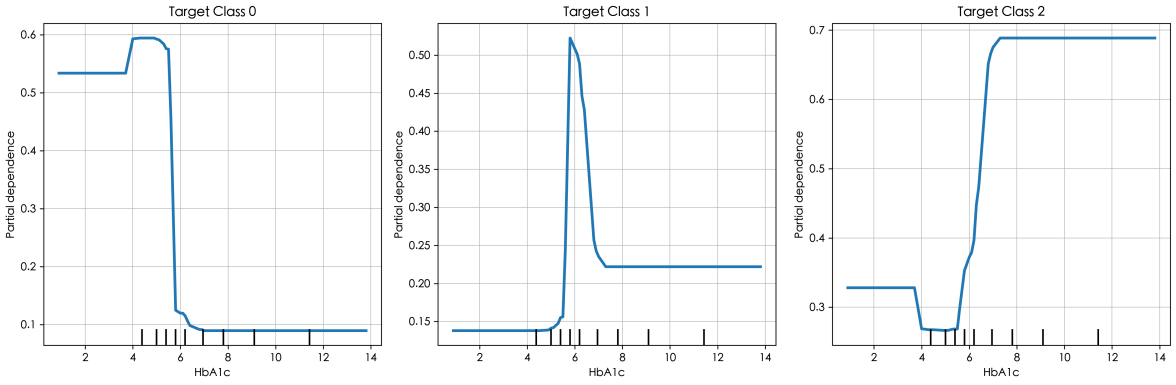

Figure 6

y 軸為屬於該類別的機率，綜合 3 張圖來觀察，可以看到 HbAlc 值在 5 以下時，被模型判斷為無糖尿病 (Class = 0) 的機率增加，HbAlc 值在 5 至 8 時，被模型判斷為糖尿病 (Class = 1) 的機率增加，在 HbAlc 值在 8 以上時，被模型判斷為糖尿病前期 (Class = 2) 的機率增加。

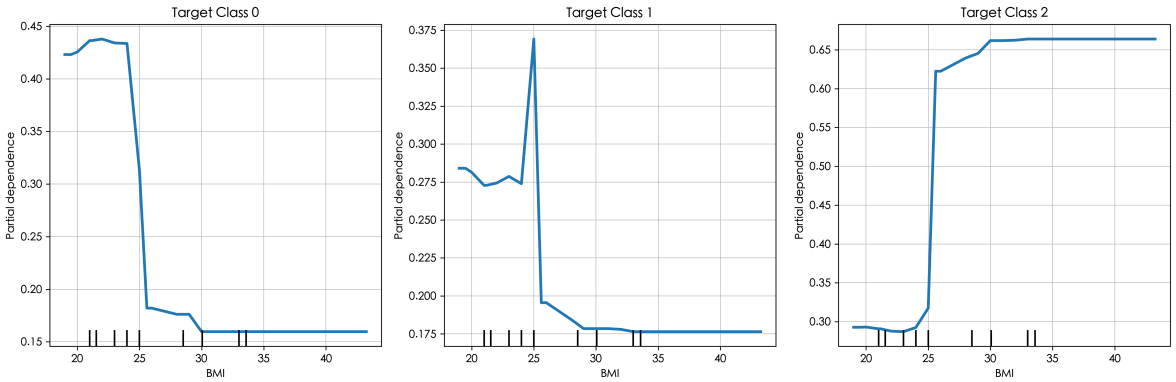

Figure 7

y 軸為屬於該類別的機率，同樣綜合 3 張圖來觀察，可以看到 BMI 值在 25 以下時，被模型判斷為無糖尿病 (Class = 0) 的機率增加，BMI 值在 25 附近時，被模型判斷為糖尿病 (Class = 1) 的機率增加，在 BMI 值在 25 以上時，被模型判斷為糖尿病前期 (Class = 2) 的機率增加。

2.2.2 XGBoost（Extreme Gradient Boosting)

模型調控參數後擬合結果模型預測的分類準確率同樣為 96%，並且畫出模型預測結果的 confusion matrix

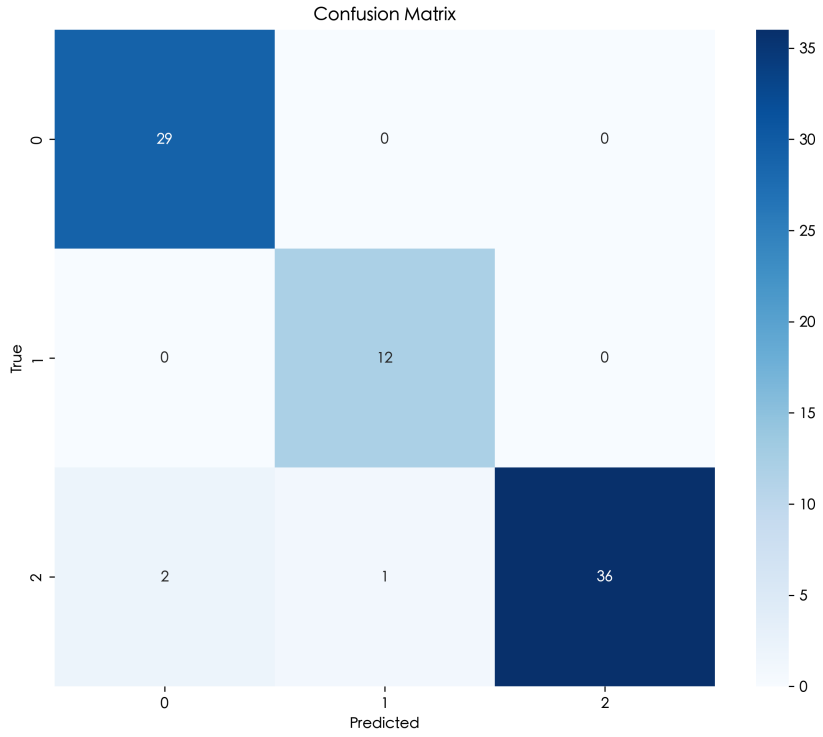

Figure 8

Xgboost 模型同樣表現良好，只有 3 個樣本預測錯誤，對於糖尿病前期 (Class = 2) 的預測準確度為 100%。接著排序出 Xgboost 模型的變數重要程度。

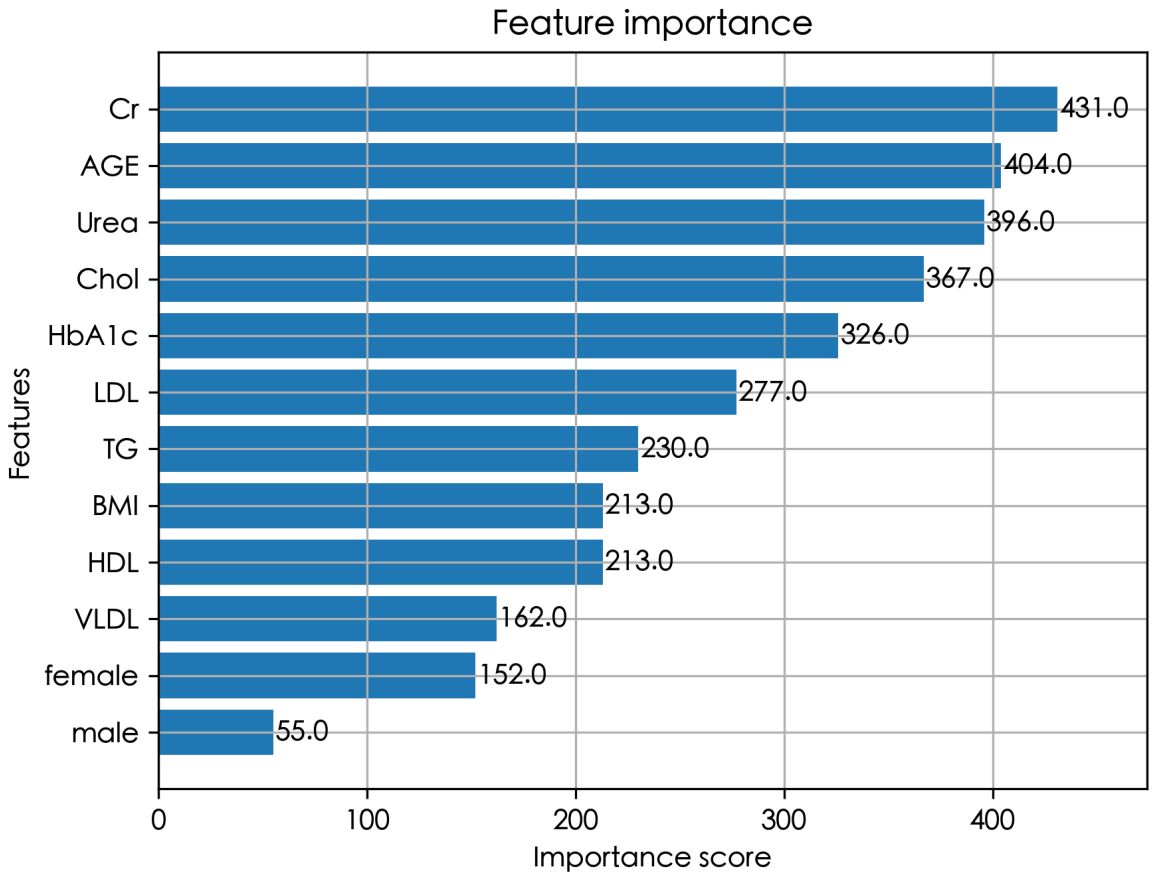

Figure 9

Random forest 模型的不同，重要的變數排序不同，並且並無展現出顯著的重要性差異。

3 預測結果與結論

根據上述模型訓練結果，可總結以下幾點結論：

1. Random forest 模型判定的前 3 名重要變數為：HbA1c，BMI，AGE，並且明顯比其他變數重要許多
2. Xgboost 模型判定的前 3 名重要變數為：Cr，AGE，Urea
3. 從散佈圖可看出糖尿病的類別預測基本上可由 2 個變數 HbA1c，BMI 分割