基於基因甲基化圖譜分析生物年齡及三高危險係數

專題編號：112-CSIE-S021
執行期限：111 年第 1 學期至 112 年第 1 學期
指導教授：白敦文
專題參與人員： 109590036 朱佑安
109590039 王泉裕
109590047 黃廷翰
109590051 陳彥宇

一、摘要

關於三高疾病（高血壓、高血糖、高血脂）與基因甲基化的關聯性尚未有任何研究報告。因此，本計劃藉由使用台灣人體生物資料庫的基因資料來分析三高患者的基因甲基化分析研究。希望能夠透過大數據分析來判定三高的基因甲基化特徵及危險程度，並結合生物年齡指標，提供一個完整精準健康評估預測系統。同時將這些資訊整合成互動式網頁介面提供給受試者使用，讓一般人可以藉由健康評估系統隨時能夠更清晰地了解自己的健康風險。

關鍵詞：表觀遺傳學、甲基化、機器學習、生物年齡、三高疾病預測

二、緣由與目的

血壓、血糖及血脂的異常升高與代謝症候群息息相關，而三高的發生常常會增加心血管疾病的風險，導致腦中風或是心肌梗塞等嚴重併發症，甚至死亡。值得關注的是，醫學診斷過程中，醫生必須綜合考慮多種指標。本研究希望透過基因甲基化重要關鍵指標的分析，結合生物年齡和三高的危險係數，提供更準確評估受試者的實際健康狀態。如此一來，受試者可早日獲悉其健康風險，並按照醫生的建議進行調整，從而更有效地預防或是控制相關代謝症候群的發生。

三、研究報告內容

(一) 研究範圍

本次研究資料來源自台灣人體生物資料庫，授權計畫編號 TWBR11012_04，是本專題指導老師白敦文教授執行計畫所申請的資料。資料蒐集時間範圍介於 2016 到 2020 年間，受試者的年齡介於 30 至 74 歲。本研究從中選取 125 位糖尿病患者、214 位高血脂患者、350 位高血壓患者以及相對沒有任何三高的 751 位健康受試者做為本次專題的研究樣本。

(二) 使用技術方法

1. XGBoost

是專為提升預測性能而設計。它結合多個弱學習器以增強模型的準確性，使用梯度提升技術，通過迭代地訓練和優化弱學習器，不斷提升整體模型的準確性。

2. 羅吉斯回歸 (Logistic Regression)

是一種用於估算資料之間二元或多類別的概率關係的統計方法，常被用於分類問題。羅吉斯回歸將輸入特徵與一個邏輯函數結合，以估計某一事件發生的概率。

3. R 語言甲基化分析

在生物信息學領域，R 語言常被用於分析高通量的生物數據，包括 DNA 甲基化資料。可以利用 minfi 和 ChAMP 進行甲基化資料的常規分析和統計建模。

4. 前端 vue+後端 flask

Vue.js 是前端 JavaScript 框架，高反

應性，可以快速地隨著數據變化而更新視圖。Flask 是一個用 Python 編寫的輕量級後端框架，它簡單且易於擴展。

(三) 架構流程

1. R 語言進行差異性甲基化位點分析

從三高患者和健康受試者中，各取 1:1 組成「糖尿病 vs 健康」、「高血壓 vs 健康」、「高血脂 vs 健康」三份資料。用 R 語言執行差異性甲基化位置分析並得到甲基化的 β 值矩陣資料。

2. 生物年齡回歸模型

先使用線性回歸初步針對「糖尿病 vs 健康」、「高血壓 vs 健康」、「高血脂 vs 健康」進行訓練，取出各自模型中前兩個到三個相對較高權重的基因位點，搭配 XGBoost 訓練出簡單且高擬合的生物年齡評估模型(圖 1)。

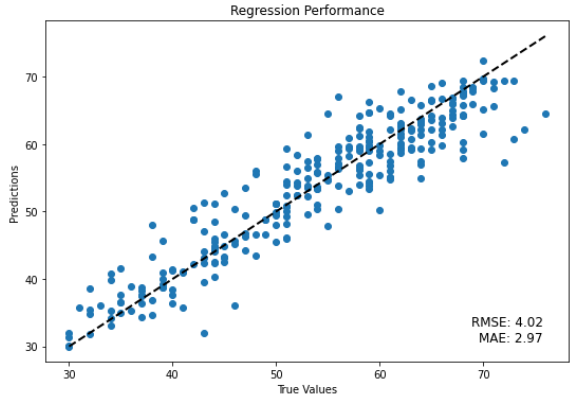

圖 1. 生物年齡模型的回歸效果

3. 三高慢性病分類模型

把「糖尿病 vs 健康」、「高血壓 vs 健康」、「高血脂 vs 健康」的 β 值轉換成 M 值，公式如(1)。各別輸出成火山圖。從火山圖挑選 $p<0.05$、差異性最大的前幾個基因位點，用來訓練羅吉斯回歸的三高慢性病預測模型。

$$M=\log_2(\beta / 1-\beta) \quad (1)$$

4. 互動式網頁呈現患者健康狀況

使用 python 前端 vue+後端 flask 撰寫，系統載入基因位點的 β 值。在後端進行生物年齡回歸和三高疾病預測，並將結果傳回前端，結合成綜合指標供使用者了解自己身體的健康狀況。

(四) 實際成果

使用者輸入自己的基因位點資料 PDF 檔，按下【start analyze】開始分析。頁面會顯示出【生物年齡/實際年齡】以及【三高各自的罹患風險】。Overall state 是用三高風險得出的綜合身體評估(圖 2)。

圖 2. 三高風險綜合健康評估系統畫面

(五) 結論

透過本專題研究，我們成功地使用甲基化的 β 值訓練生物年齡的回歸模型以及三高疾病的分類預測模型。這些模型能夠評估個體的健康狀態和預測潛在的疾病風險。更進一步可以將這些模型整合進一個互動式網頁，使用者能夠輕鬆地查看自己的身體健康狀況。

然而，如同所有的預測工具一樣，我們的模型仍有其局限性。希望未來能夠收集更多的資料以提高模型的準確性。

此外，也希望能夠擴展網頁的進階功能，例如提供個性化的健康建議或連結專業的健保醫療資源，使得用戶不僅能了解自己的健康狀態，還能夠採取具體的行動以改善或維護其健康。

四、參考文獻

Ziwei Ye, Lirong Jiang, Mengyao Zhao , Jing Liu, Hao Dai, Yiping Hou, and Zheng Wang, “Epigenome-wide screening of CpG markers to develop a multiplex methylation SNaPshot assay for age prediction”, *Legal Medicine*, 102-115, 2022.