中華民國第 63 屆中小學科學展覽會

作品說明書

高中組　電腦與資訊學科

052512

以大腸直腸癌預測為例進行缺失值處理方式的探討與實驗

學校名稱：國立宜蘭高級中學

作者：	指導老師：
高二 林紹群	梁祐銘

關鍵詞：機器學習、大數據、大腸直腸癌

摘要

機器學習和精準醫療是目前醫學界的熱門話題。機器學習在醫療領域的應用越來越普及，可幫助臨床更快速及精準診斷疾病，並提供個人化治療方案。例如，通過訓練大量醫學影像數據，建立深度學習模型，可用於腫瘤的自動辨識與分類。通過醫療資料大數據分析，可以為臨床提供及時的疾病預測和預防建議。然而，如何讓臨床資料結合機器學習建立模型預測，是很重要的議題。本研究使用臺北醫學大學數據處蒐集衛生福利部雙和醫院的大腸直腸癌與大腸炎病患三年的臨床資料，結合機器學習進行模型的建立與預測。經處理數據的缺失值、特徵的排序與選取及向前特徵選取法來訓練與驗證模型，找出分辨大腸直腸癌和大腸炎的最佳檢驗項目組合及效能，以預測大腸直腸癌。

壹、前言

一、研究動機

大腸直腸癌(Colorectal cancer, CRC)現在的主要篩檢方式，是透過糞便潛血檢測的方式進行檢測。而現在的糞便潛血的檢測方式主要可以分為二種，包含 Gauaiac-based FOBTs (gFOBT)與 Immunochemical FOBTs (iFOBT)。gFOBT 的檢測方式是透過氧化還原的方式對糞便當中的血紅素進行檢測，然而 gFOBT 也會對其他能夠進行氧化還原反應的成分呈現陽性，譬如其他動物的血紅素、天然的氧化劑等。因此，現在臨床上大多使用的都是 iFOBT，其檢測方式是透過免疫的方式，利用抗體辨識人類的血紅素來檢測糞便中是否潛血。這二種檢測在一個系統性回顧研究當中被一起比較。gFOBT：靈敏度 62-79%與特異性 87-96%; iFOBT：靈敏度 79%與特異性 94%[1]。儘管這二種糞便潛血的方式都擁有不錯的靈敏度(Sensitivity)與特異性(Specificity)，但因為糞便採集送檢的過程並非一次到院就能夠完成，而是需要病患將採檢管攜帶回家後採檢完畢，再將檢體帶回醫院進行檢查。因此，願意進行糞便潛血檢測的病患將大幅度減少。

在科技部補助的專題研究報告『台灣地區民眾大腸直腸癌篩檢行為與行為誘因之關聯性：調查與介入研究』中提到：民國 93 年的糞便篩檢涵蓋率為 4.8%，而在推動之後在民國 102 年已經達到 38.2%。相較於其他癌症的篩檢，如乳房攝影篩檢較困難推動，主要的原因為：糞便篩檢程序較複雜，民眾參與的意願較低。此外，因為篩檢管需要發放後再繳回，使得民眾的配合意願較低。醫院診所也需要為了衛教民眾篩檢管的使用方式，佔據部分人力資源，有較高的服務成本。然而，有四成的民眾為糞便潛血陽性，卻不願意進一步的進行大腸直腸鏡檢查，使得陽性個案的後續追蹤不易[2]。

相較之下，若能夠讓民眾在醫院或診所進行抽血健檢時，就能夠完成大腸直腸癌的篩檢的話，必定能夠提升大腸直腸癌篩檢的涵蓋率，並且提升後續陽性個案的追蹤。臨床現有與大腸直腸癌相關血液檢測項目為癌胚抗原(Carcubienbryibuc antigen, CEA)，但不能夠做為早期診斷大腸直腸癌的工具，而是一個癒後指標(Prognostic biomarker)。如同新藥開發過程費時長、成本高，在開發新的血液檢測項目也同樣的需要許多的時間、人力與資源。我考完國中會考後就開始學習 Python，並於去年開始學習使用臨床數據結合機器學習進行模型(Model)的建立與預測，加上已有許多研究試著將臨床醫院的資料應用到對病人的檢測上。因此，我決定要使用 Python 建立機器學習模型並且用以偵測患者是否罹患大腸直腸癌。

二、目的

主要目的: 健康檢查的血液與尿液檢測項目做出機器學習預測模型

(一) 哪一種補值的方法最合適?

(二) 哪一些檢測項目最適合用於預測大腸直腸癌?

(三) 是否可以透過處理缺失值、特徵選擇加強機器學習的表現?

三、文獻回顧

(一) 機器學習的訓練、驗證與預測

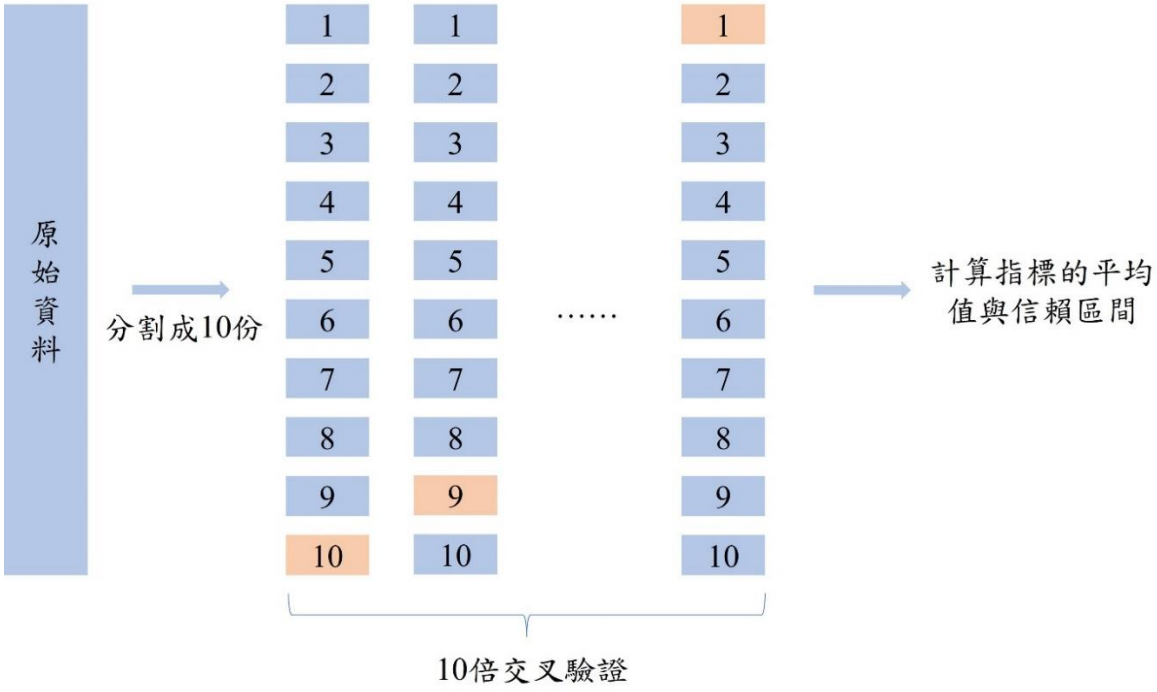
1. 機器學習及其歷史

人工智慧、機器學習與深度學習是容易令人混淆的專有名詞。機器學習(Machine Learning)最早於 1959 年被提出，指的是能訓練電腦以數學或特定模式辨認的方式進行學習。而人工智慧(Artificial intellegence)則是泛指任何能夠展現出人類智慧的電腦活動，例如使用簡單的 IF 判斷式進行判斷。深度學習(Deep learning)則是指以類神經網絡(Neural network)所建立的模型，因其透過多個不同層(Layer)的函數間傳遞變數，類似神經元之間的衝動，故稱為深度學習[3]。

2. 機器學習的運作模式

機器學習的訓練(Train)與函數(Function)的 Fitting 相似，藉由資料與欲預測之結果帶入機器學習的函數或邏輯，再藉由計算並調整函數或邏輯當中的參數，使其符合一定應對關係，達到藉已知事物之關係推測未知事物之關係。機器學習的訓練需要龐大的資料，訓練後更需要用新的資料進行機器學習後的驗證。但資料的取得不容易，因此在訓練與驗證的方式上，蔚為風行的是 X 倍交叉驗證(X-fold cross validation)方式，即將資料隨機切分為 X 份，使用 X-1/X 的資料進行訓練，並用 1/X 的資料進行驗證，訓練與驗證進行 X 次，最後再將驗證得到的分數相加後除以 X，得到平均的驗證結果[4]。本研究使用 10 倍交叉驗證(圖一)

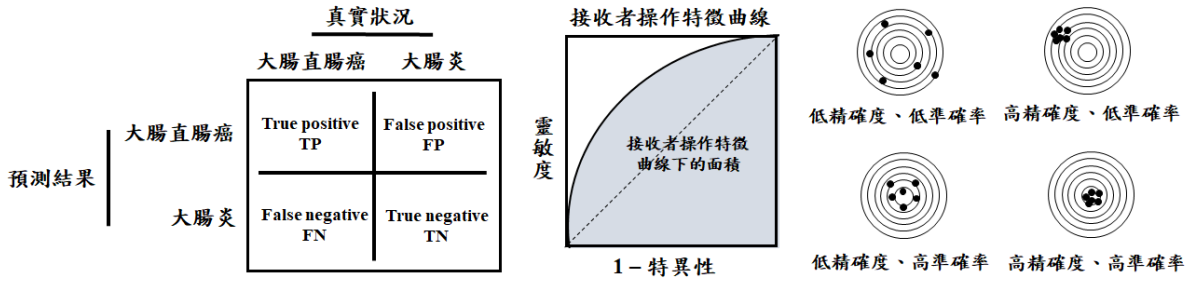
。



圖一、10 倍交叉驗證示意圖

3. 評估機器學習的成果

機器學習的訓練與驗證的成果，可以混淆矩陣(Confusion matrix)表示(圖二)，包含了真陽性(True positive)、真陰性(True negative)、偽陽性(False positive)、偽陰性(False negative)。混淆矩陣能夠進一步的計算出準確度(Accuracy)、精確度(Precision)、靈敏度及特異性[5]。『靈敏度』與『1-特異性』的數據可以進一步繪出接收者操作特徵曲線(Receiver operating characteristic curve, ROC curve)，接收者操作特徵曲線下面積(Area under the ROC curve, AUC)更是全面評估機器學習效能的指標，接收者操作特徵曲線下面積值在 0.5 和 1 之間。接收者操作特徵曲線下面積值越接近 1.0，預測越接近真實性[6]。



$$\text{準確率} = \text{人群(大腸直腸癌+大腸炎)中，準確預測是大腸直腸癌的比率} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{靈敏度} = \text{罹患大腸直腸癌患者，預測是大腸直腸癌的比率} = \frac{TP}{TP+FN}$$

$$\text{特異性} = \text{罹患大腸炎患者，預測是大腸炎的比率} = \frac{TN}{FP+TN}$$

$$\text{精確度} = \text{預測是大腸直腸癌患者，診斷試驗結果是大腸直腸癌的比率又稱陽性預測值} = \frac{TP}{TP+FP}$$

圖二、混淆矩陣

4. 機器學習的種類

機器學習可以學習的模式分為兩大類: 監督式學習(Supervised learning)與非監督式學習(Non-supervised learning)。監督式學習指的是在機器學習的訓練與驗證的過程中，會在每一筆資料上給予標籤(label)，譬如一連串的檢驗項目是屬於大腸直腸癌病患或者健康人，又或者是X光片是屬於肺炎患者或肺癌患者。監督式學習時，機器學習會將會學習如何藉由資料上的特徵去區分不同的標籤[7]。非監督式學習指的是機器學習的訓練與驗證過程中，在每一筆資料上不會給予標籤，因此機器學習會將特徵相近的資料連結起來，譬如他可能會將身高相似的人當成一個群體、體重相近的人當成一個群體。非監督式學習在單細胞基因體學或是腸道菌叢的研究使用得較為廣泛，將單細胞或腸道菌叢的表達相似的基因連結在一起，發現健康人和癌症患者被歸類在同一個群組，故可能代表健康人也有罹癌風險[8]。

5. 常用於機器學習的模型及其特點

常用於機器學習的演算法(Algorithms)包含決策樹(Decision tree)、隨機森林(Random forest)、支援向量機(Support vector machine, SVM)、極限梯度提升(eXtreme gradient boosting, XGBoost)、輕量化梯度提升機(Light gradient boosting machine, LightGBM)與邏輯式回歸(Logistic regression)。

(1) 決策樹

決策樹的學習方式是藉由計算資料分類後群組(Group)的亂度，再以分類後亂度最小的當作最佳選擇。在決定一次最佳選擇後，決策樹會產生一個分支(branch)，再重複進行學習，找出分類後亂度最小的項目及其閥值(Cut-off value)，再產生另一個分支。如此一直升長下去，直到亂度為零[9]。

(2) 隨機森林

隨機森林的學習方式，是藉由重複抽取大筆資料中的隨機小筆資料進行一個決策樹的學習，接著再重複這個過程直到建立出許多的決策樹。在針對資料進行分類時，經由學習過程建立出個每一個決策樹都會做出預測，最終再以多數決的方式，做出隨機森林對該筆資料的分類結果[10]。

(3) 支援向量機

支援向量機是一種常見的監督式機器學習演算法，經常被用於解決分類問題。支援向量機的分類方式，是透過找到一個能夠最大化兩個類別(Class)之間的高維度平面，藉此區隔兩個不同類別，達到分類的目的。他的優勢是可以處理高維度的數據，並且可以有效解決非線性分類的問題[11]。

(4) 極限梯度提升

XGBoost 是由美國華盛頓大學博士生陳天奇所提出，是以 Gradient Boosting 為基礎進行的改良。Boosting 是指將複雜程度低、學習效能差的的模型們綁在一起，再以多數決的方式達成最終分類結果。XGBoost 的改良包含了: 利用局部近似算法對分類節點的最佳化、損失函數中加入 L1/L2 以控制模型複雜度、以及提供 GPU 平行化運算[12]。

(5) 輕量化梯度提升機

LightGBM 也是屬於一種 Gradient boosting，是由微軟團隊所提出。這個模型是基於 XGBoost 的缺點再次進行改良，是較輕量化的模型。在官方的文件中說明了 LightGBM 的優點: 更快的訓練速度和更高的效率、低記憶體使用、能夠處理大規模數據[13]。

(6) 邏輯式回歸

邏輯式回歸是一種二元或多元的分類演算法，是探討依變數(Y)與自變數(X)的關係，可以用於準確預測。依變數是類別變數，可以是『有或無』、『是或否』及『同意或不同意』。自變數是二元或多元的連續變數或類別變數[14]。

(二) 機器學習與特徵工程

1. 機器學習的侷限性之一

機器學習並非萬能，其中一個重要的局限性是來自於訓練機器學習時所使用的特徵(Feature)。特徵是資料當中的項目，譬如糖尿病患者的血糖即屬於特徵，因為患者的血糖數值具有預測糖尿病的特徵。然而，有時錯誤或資訊量較低的特徵反而會成為機器學習時的干擾。舉例來說，與糖尿病的發生較無相關性的甲狀腺激素，便無法使得機器學習的預測更加準確[15, 16]。

2. 特徵工程的概念: 特徵挑選或刪除

處理特徵的知識統稱為特徵工程(Feature engineering)，主要目的便是篩選出有用的特徵，或者將無效的特徵刪除掉[17]。總而言之，就是經由各類方式使資料去蕪存菁。有趣的是，並非所有的機器學習模型都需要使用到特徵工程，譬如深度學習的特點之一就是他解決了特徵的干擾，因為在不同層間傳遞函數結果的同時，也起到了特徵選取的功用。

3. 常見的特徵工程做法

特徵工程的常見做法有三類，包含特徵選取、特徵刪除、以及創造特徵。特徵選取(Feature selection)，是使用統計方法或關聯性分析，將對分類最有貢獻的特徵挑選出來並且排序。特徵刪除(Feature remove)，是在訓練與驗證的過程中，藉由將特徵移除來提升機器學習的效能。最後則是創造特徵(Feature creation)，指的是將某些特徵的數值先合併運算，來達到創建新特徵的效能。腎絲球過濾率估計值(Estimated glomerular filtration

rate, eGFR)是用來評估腎損傷的指標，就是透過合併血清肌酸酐、年齡與性別所創造出的新特徵[18, 19]。

（三）醫院檢驗項目與機器學習的應用

醫院的資料現在已經被視為是一座寶貴的大數據金庫，已經被許多科學家先後的開採以及建立了機器學習的模型並發展成應用。許多科學家也指出，使用機器學習與醫院的檢驗項目來建立應用程式是具有潛力的[20]。如 Olmedo 等人便利用了醫院的臨床檢驗項目來預測病患感染 COVID-19 後的嚴重程度與死亡率[21]。如 Bai 等人利用了常規的實驗室檢驗項目來達到預測慢性腎臟病病患是否具有發展成腎衰竭的風險[22]。除了感染性疾病與慢性疾病外，機器學習也被應用在協助診斷或偵測癌症上。如 Gould 等人利用常規臨床資料和檢驗項目的資料來達到早期偵測肺癌的應用[23]。

前人已有許多利用機器學習和醫院的檢驗項目資料來發展應用並達到針對疾病的早期預測的研究，代表著這類型研究的可行性很高。因此本次的研究目的是要以常規檢驗項目做出能夠預測大腸直腸癌的機器學習模型。

貳、研究設備與器材

一、設備

筆記型電腦。

二、軟體

Python、Weka、Scikit learn、Anaconda3 (Jupyter Notebook)與 Excel。

三、大數據

臨床實驗室檢驗項目的數據取得，獲得**臺北醫學大學暨附屬醫院聯合人體研究倫理委員會**(N202201049)的核准，臨床數據源自**衛生福利部雙和醫院**(2018/01/01 至 2020/12/31)。

參、研究過程與方法

一、資料處理

臨床數據選取的規定是以同一個病例號的每一項檢驗項目，取得最新一筆數據。原始數據經過資料前處理，本研究取得臨床數據 8047 筆資料其含有缺失值的檔案(.csv)並開始進行後續研究。本研究只**針對大腸直腸癌與大腸炎**(Colitis)患者進行數據收集及分析。**缺失值的定義是臨床檢驗項目(特徵)沒有數據**，代表病人在(2018/01/01 至 2020/12/31)期間沒有進行此項目的檢測。至於有檢測沒輸入數據或數入錯誤數據的失誤，會由醫院檢驗科均有建置的實驗室資訊管理系統(Laboratory information management system, LIMS)來預防失誤事件。

二、缺失值處理

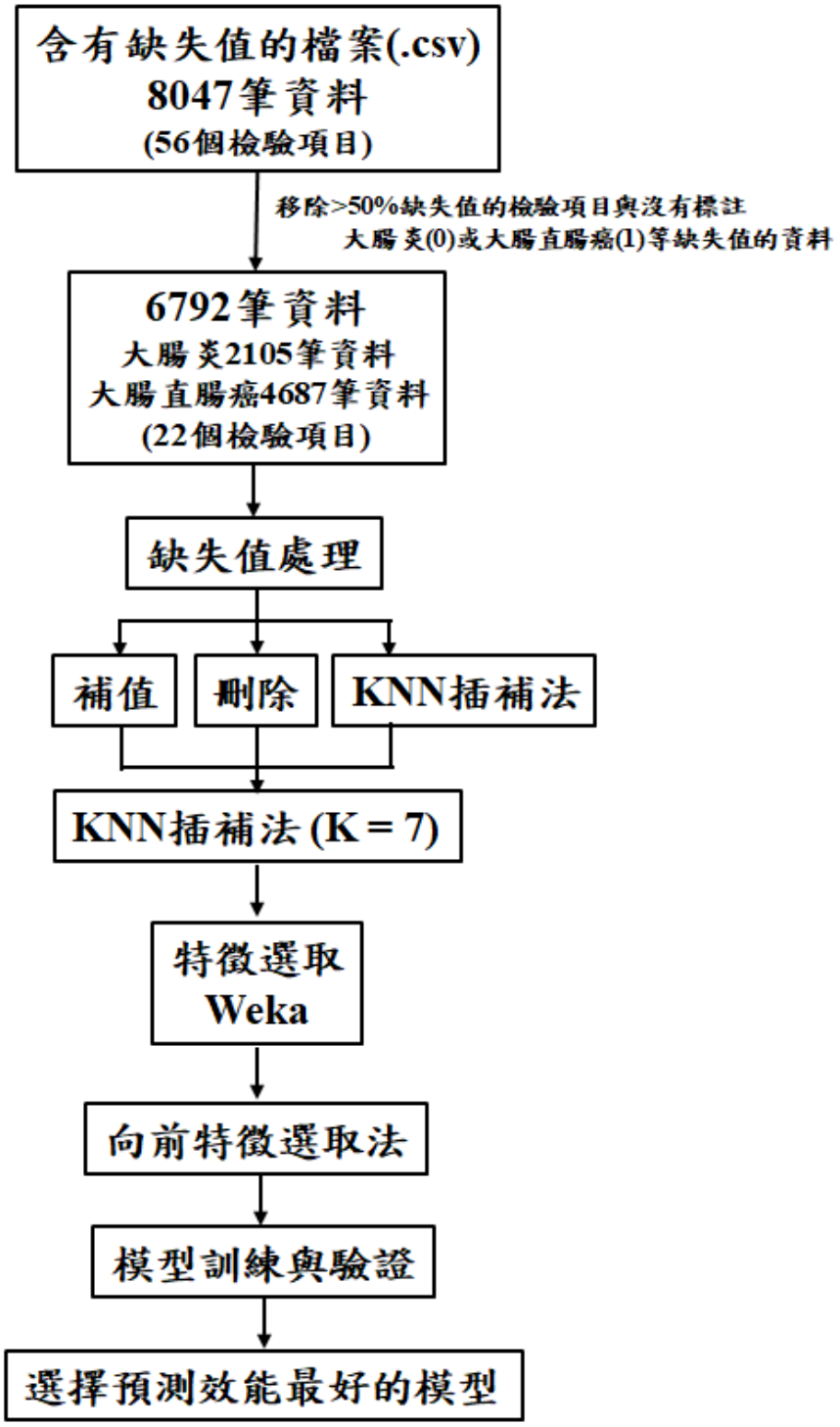
從臨床實驗室檢驗項目中，搜集了共 56 個檢驗項目結果。若檢驗項目的缺失比例達整體病人的 50%，則將該檢驗項目移除。處理過後，剩餘 22 個檢驗項目，其缺失比例為 0 到 44.1% (表一)。缺失資料的處理方式以補值、刪除及 KNN 插補法(KNN Imputation)進行補值。補值是以平均值(連續性變項)及中位數(類別變項)填補。刪除是以若病人有任一缺失，則將該病人移除。KNN 插補法進行補值則是利用無缺失的鄰近值，計算缺失的數值。先以學生 t 檢定的方式比較原始資料與填補資料是否達到統計上的顯著差異，再由邏輯式回歸訓練與驗證之準確度的標準差結果進行比較。

三、特徵選取

使用 Weka 3.8.3 進行特徵排序(Feature ranking)。

四、模型訓練與驗證

本研究以**向前特徵選取法**(Forward feature selection method)來進行特徵選取，並使用 10 倍交叉驗證將實驗數據分成訓練資料集與驗證資料集。使用 Scikit learn 0.21.3 框架中使用 Decision tree、Random forest、SVM、XGBoost、LightGBM 與 Logistic regression，每一個模型都經過測試以找到最適合的參數條件使用。參數測試在 decision tree 中測試了 kernel (gini, entropy)與 depth (1 ~ 10)，而在 Random forest 中測試了 kernel (gini, entropy)與 tree number (100 ~ 500)。在 SVM 中，測試了 gamma 值(1e-6 ~ 1e-10)與 C 值(1e5 ~ 1e7)。在 XGBoost 中測試了 eta 值(0.01 ~ 0.2)與 depth (1 ~ 10)，然後在 LightGBM 中測試了 leaves (50 ~ 400)與 depth (1 ~ 10)。Logistic regression 不需設定參數。研究經反覆測試，利用準確度作為進步依據進行模型的調整，並且最終將預測正確與錯誤的值整合，以準確度(Accuracy)、精確度(Precision)、靈敏度(Sensitivity)、特異性(Specificity)與接收者操作特徵曲線下面積(AUC)來呈現模型評估結果。圖三是研究過程與方法的流程圖。



圖三、研究過程與方法的流程圖

肆、研究結果

一、資料處理[實驗一]

經由衛生福利部雙和醫院蒐集資料，總共取得 24863 筆資料。整理後發現共有 8047 筆資料，隨機抽取資料繪出缺失值的視覺化圖(圖四)。本次研究只取用檢驗項目的資訊，移除>50%缺失值的檢驗項目與沒有標註大腸炎(0)或大腸直腸癌(1)等缺失值的資料(如圖四中第 1647 筆資料)，最終剩下 6792 筆資料，其中大腸炎有 2105 筆資料、大腸直腸癌有 4687 筆資料。數據重複檢查，經每一筆資料的驗證確認大腸炎與大腸直腸癌沒有雙重登錄。大腸炎與大腸直腸癌的樣本數比例是 1:2.23。

二、缺失值處理[實驗一]

圖四、缺失值的視覺化圖

缺失值由圖四可以看出，某些項目的缺失值是一起缺失的，例如沒有 eGFR 的病患也沒有 Creatinine。Platelet、Neutrophil、MCH、MCHC、RBC、Monocyte、Eosinophil、MCV、WBC、HCT、HGB、Basophil 和 Lymphocyte 等血球項目也試共同缺失。

針對缺失值的部分，選用了三種不同的處理方式，包含刪除(Delete CRC)、使用平均值填值(Mean CRC)、以及 KNN 補值方式 (K2-K10)。補值方法的效能則先以統計檢定的方式，針對原始資料與補值後的資料進行比較。表一以平均值±標準差的方式表現，以刪除的方式處理缺失值，刪除後的資料與原始資料的資料型態有 15 個項目呈現統計上的顯著差異，包含 eGFR、Creatinine、Platelet、MCH、MCHC、RBC、Eosinophil、MCV、HCT、HGB、K、BUN、CRP、RDW-CV 和 Random Glucose。平均值補值的方式則沒有呈現任何的顯著差異。使用 KNN 補值方式的資料，則皆有 4 個顯著差異的項目，包含 BUN、CRP、RDW-CV 和 Random Glucose。

若以統計檢定的結果最為判斷，會認為平均值可以得到最好的補值資料。然而，所使用的統計檢定是學生 t 檢定，他的統計檢定方法主要便是利用比較兩筆資料的平均值變化來進行判斷，因此使用平均值填補資料後，和原始資料相比，他們的平均值不會產生變化。因此，利用了以邏輯式回歸效能的準確度的平均值和標準差當作評量方式(圖五)。單獨的使用了每份資料中的每一個特徵進行多次訓練，並且以訓練結果的標準差做為參考指標。當訓練結果的標準差越小，就表示該資料越能夠使模型的區分擁有較小的誤差。根據圖五所示，不同項目單獨訓練的標準差累積值最小的是使用 KNN 補值法且 K=7，因此後續的模型訓練與驗證會使用 KNN 補值法且 K=7 的資料進行。

表一、原始資料與補值後資料的比較

	Origin CRC	Delete CRC	Mean CRC	K2	K3
eGFR	**135.63±157.93**	117.08±146.22 **	135.63±139.36	135.84±140.77	135.67±140.41
Creatinine	**1±1.14**	1.3±1.66 **	1±1.01	0.99±1.01	0.99±1.01
Platelet	**241.08±90.96**	235.5±107.25 *	241.11±78.36	242.1±80.04	242.29±79.59
Neutrophil	**69.48±15.61**	69.39±15.91	69.48±13.43	69.34±13.84	69.32±13.71
MCH	**29.47±3.36**	29.75±3.46 *	29.47±2.9	29.47±2.94	29.47±2.93
MCHC	**34.01±1.02**	33.94±1.07 *	34.01±0.88	34.03±0.9	34.03±0.89
RBC	**4.48±0.8**	4.18±0.89 **	4.48±0.69	4.5±0.7	4.5±0.7
Monocyte	**7.76±3.65**	7.83±4.04	7.76±3.14	7.76±3.2	7.77±3.18
Eosinophil	**1.69±2.11**	1.9±2.39 *	1.69±1.81	1.69±1.85	1.69±1.84
MCV	**86.53±8.69**	87.52±8.94 **	86.53±7.49	86.5±7.61	86.5±7.57
WBC	**9.51±4.9**	9.53±5.93	9.51±4.22	9.49±4.29	9.49±4.26
HCT	**38.5±6.26**	36.23±7.03 **	38.49±5.4	38.65±5.49	38.66±5.47
HGB	**13.11±2.22**	12.31±2.46 **	13.11±1.91	13.16±1.95	13.17±1.94
Basophil	**0.48±0.39**	0.49±0.42	0.48±0.33	0.48±0.34	0.48±0.34
Lymphocyte	**20.27±13.1**	19.86±13.32	20.27±11.27	20.42±11.64	20.44±11.53
K	**3.82±0.54**	3.86±0.64 *	3.82±0.45	3.82±0.47	3.82±0.46
BUN	**18.69±19.26**	21.97±23.45 **	18.69±14.2	17.07±14.6 **	17.07±14.58 **
Na	**137.75±3.77**	137.65±4.67	137.75±3.2	137.82±3.26	137.83±3.24
GOT	**36.54±215.37**	46.6±356.7	36.54±184.98	35.98±185.22	35.97±185.16
CRP	**3.66±5.69**	4.41±6.56 **	3.66±4.22	3.29±4.49 *	3.31±4.42 *
RDW-CV	**14.61±2.55**	15.18±2.91 **	14.61±1.86	14.4±2.03 **	14.4±2 **
Random Glucose	**130.05±60.76**	142.67±74.46 **	130.05±48.04	125.63±50.1 **	125.79±49.69 **

*** $p < 0.05$, ** $p < 0.001$**

延續表一

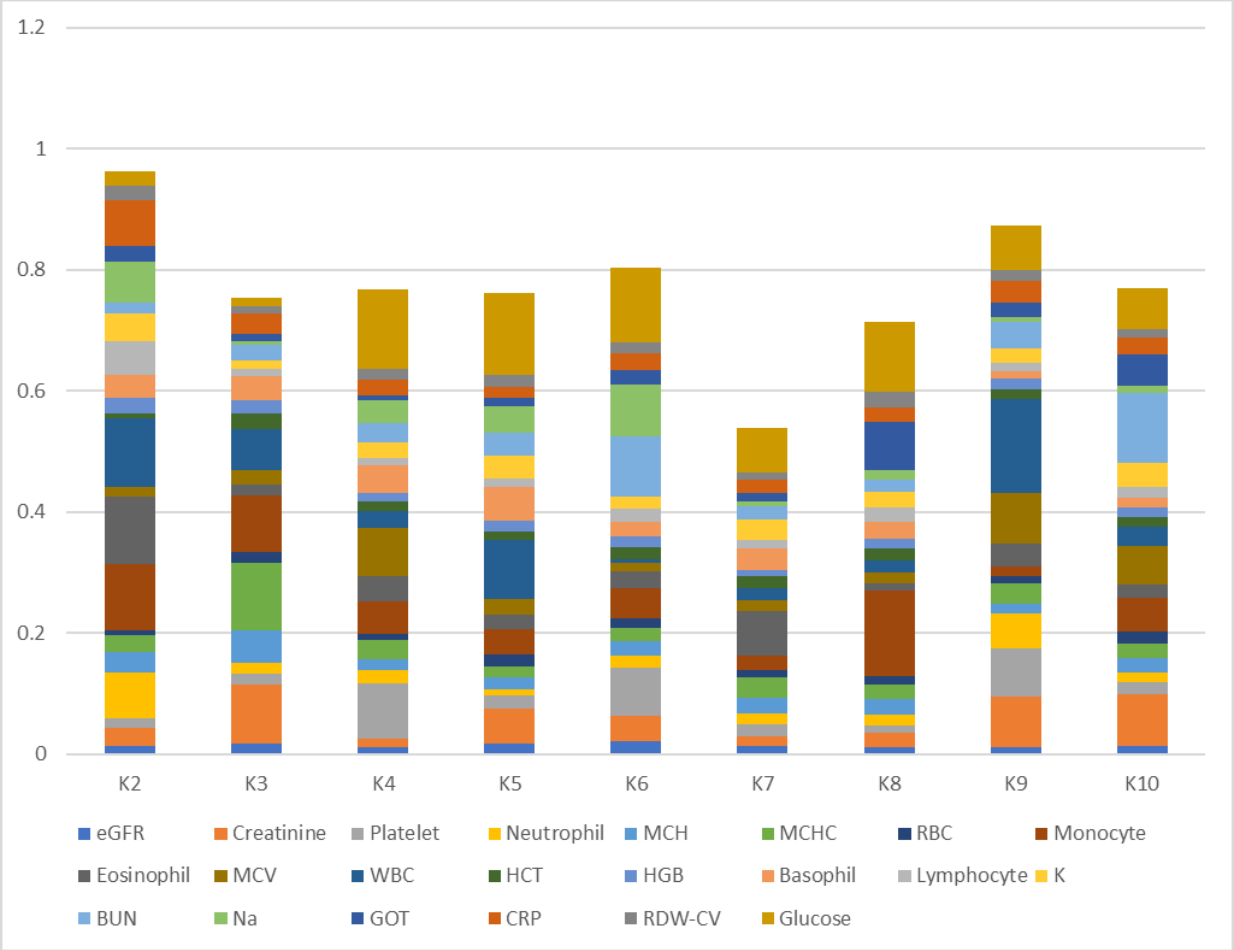
	K4	K5	K6	K7	K8
eGFR	136±140.93	136.13±140.76	136.1±140.48	136.09±140.36	136.02±140.23
Creatinine	0.99±1.01	0.99±1.01	0.99±1.01	0.99±1.01	0.99±1.01
Platelet	242.11±79.27	242.14±79.08	242.25±79.05	242.22±78.98	242.19±78.93
Neutrophil	69.26±13.68	69.27±13.62	69.27±13.6	69.28±13.57	69.28±13.56
MCH	29.48±2.93	29.48±2.92	29.48±2.92	29.48±2.91	29.48±2.91
MCHC	34.03±0.89	34.03±0.89	34.03±0.89	34.03±0.89	34.03±0.88
RBC	4.5±0.69	4.5±0.69	4.5±0.69	4.5±0.69	4.5±0.69
Monocyte	7.77±3.17	7.76±3.17	7.76±3.16	7.76±3.16	7.76±3.16
Eosinophil	1.69±1.83	1.69±1.83	1.69±1.83	1.69±1.82	1.69±1.82
MCV	86.52±7.56	86.52±7.55	86.51±7.54	86.51±7.53	86.52±7.53
WBC	9.49±4.26	9.49±4.25	9.5±4.24	9.5±4.24	9.5±4.24
HCT	38.67±5.47	38.67±5.46	38.67±5.46	38.67±5.45	38.66±5.45
HGB	13.17±1.94	13.17±1.94	13.17±1.94	13.17±1.93	13.17±1.93
Basophil	0.48±0.34	0.48±0.34	0.48±0.34	0.48±0.34	0.48±0.34
Lymphocyte	20.5±11.49	20.5±11.45	20.5±11.42	20.49±11.4	20.49±11.39
K	3.82±0.46	3.82±0.46	3.82±0.46	3.82±0.46	3.82±0.46
BUN	17.07±14.56 **	17.06±14.54 **	17.06±14.54 **	17.06±14.55 **	17.06±14.55 **
Na	137.82±3.23	137.82±3.23	137.82±3.22	137.82±3.22	137.81±3.22
GOT	35.91±185.06	35.89±185.04	35.87±185.03	35.86±185.02	35.86±185.02
CRP	3.3±4.37 *	3.29±4.35 *	3.29±4.34 *	3.29±4.32 *	3.3±4.32 *
RDW-CV	14.4±1.98 **	14.4±1.97 **	14.4±1.96 **	14.4±1.96 **	14.41±1.96 **
Random Glucose	125.98±49.45 **	126.26±49.28 *	126.45±49.15 *	126.62±49.06 *	126.73±48.98 *

*** $p < 0.05$, ** $p < 0.001$**

延續表一

	K9	K10
eGFR	136.04±140.17	136.06±140.17
Creatinine	0.99±1.01	0.99±1.01
Platelet	242.22±78.89	242.18±78.84
Neutrophil	69.3±13.55	69.3±13.54
MCH	29.48±2.91	29.47±2.91
MCHC	34.02±0.88	34.02±0.88
RBC	4.5±0.69	4.5±0.69
Monocyte	7.76±3.16	7.76±3.16
Eosinophil	1.69±1.82	1.69±1.82
MCV	86.51±7.52	86.51±7.52
WBC	9.5±4.24	9.5±4.24
HCT	38.66±5.45	38.66±5.45
HGB	13.17±1.93	13.17±1.93
Basophil	0.48±0.34	0.48±0.34
Lymphocyte	20.48±11.38	20.47±11.37
K	3.82±0.46	3.82±0.46
BUN	17.06±14.55 **	17.06±14.55 **
Na	137.81±3.22	137.81±3.22
GOT	35.86±185.02	35.85±185.02
CRP	3.3±4.31 *	3.31±4.31 *
RDW-CV	14.41±1.95 **	14.41±1.95 **
Random Glucose	126.83±48.91 *	126.94±48.84 *

*** $p < 0.05$, ** $p < 0.001$**



圖五、以邏輯式回歸模型進行訓練時得到的標準差累積結果

三、特徵選取[實驗二]

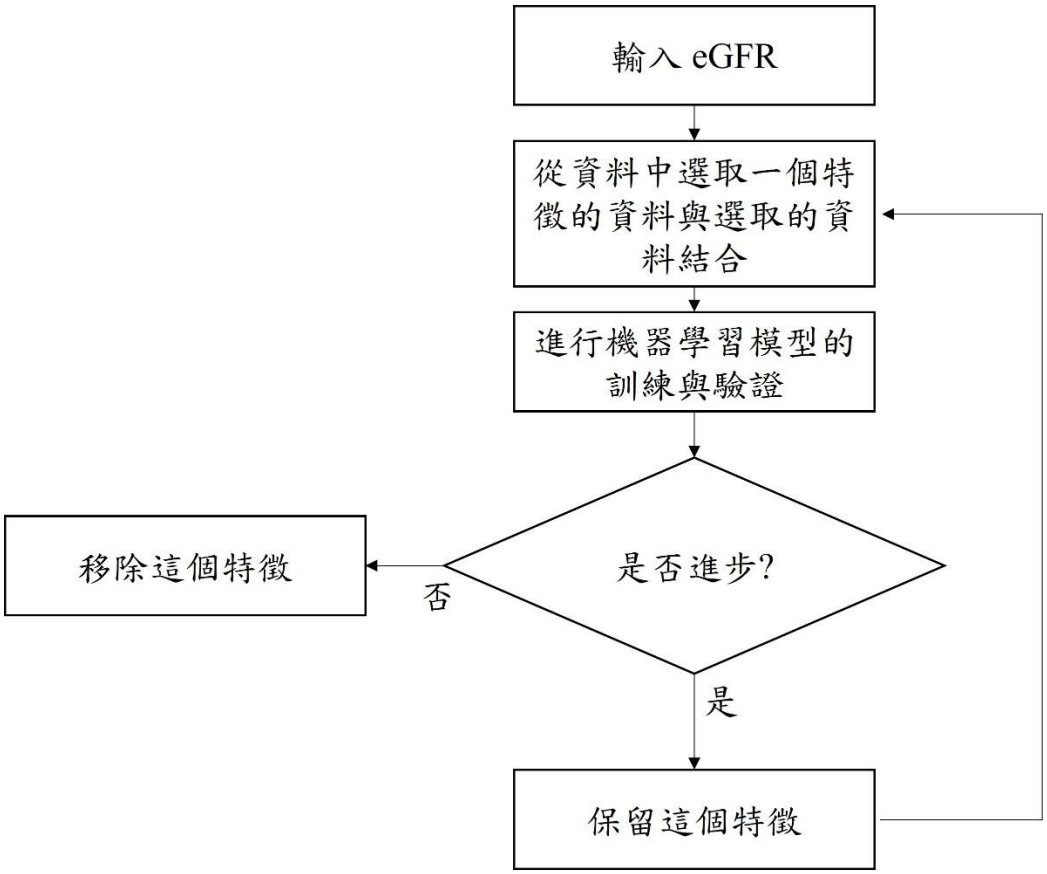
利用 Weka 進行特徵排序。使用 InfoGainAttributeEval (InfoGain) + Ranker 的方式進行特徵排序，特徵排序的結果如表二所示。由高至低，依序為 eGFR、RBC、BUN、RDW-CV、HGB、HCT、Creatinine、MCV、MCH、Platelet、WBC、CRP、Na、K、Eosinophil、Basophil、Random Glucose、MCHC、GOT、Lymphocyte、Monocyte 與 Neutrophil。

表二、使用 Weka 進行特徵排序

Ranking Score	Attributes
0.0894	**eGFR**
0.0832	RBC
0.0821	BUN
0.0748	RDW-CV
0.0643	HGB
0.0613	HCT
0.0572	Creatinine
0.0443	MCV
0.0435	MCH
0.0431	Platelet
0.042	WBC
0.0353	CRP
0.0347	Na
0.0313	K
0.028	Eosinophil
0.0277	Basophil
0.0269	Random Glucose
0.0222	MCHC
0.0221	GOT
0.0209	Lymphocyte
0.0195	Monocyte
0.0162	Neutrophil

四、模型訓練與驗證[實驗三]

使用**向前特徵選取法**，以 eGFR 為起始特徵，並且隨機的選取新的特徵，若特徵加入後能夠使得模型的評分增加，就保留這個特徵並且選取一個新的特徵。若不能，就移除這個特徵，再選取新的特徵，**以較佳的準確度判定進步**(圖六)。利用準確度檢視整體正確率，利用精確度檢視針對陽性個體的正確率，利用靈敏度檢視陽性個體的檢出率,利用特異性檢視陰性個體的正確率，利用接收者操作特徵曲線下面積來檢視靈敏度與特異性的綜合表現。在模型的訓練與驗證中發現訓練驗證的結果為精確度低但是準確度>70%，大腸炎與大腸直腸癌的樣本數比例是 1:2.23，大腸直腸癌的樣本數過多，顯然是分類結果異常，因此以 Imblearn 進行調整。調整後得準確度(Accuracy) 79.1%-80.3%、精確度(Precision) 78.7%-83.5%、靈敏度(Sensitivity) 91.5%-98.6%、特異性(Specificity) 7.8%-37.9%及接收者操作特徵曲線下面積(AUC) 0.76-0.81。其中，表現得最好的模型為 LightGBM，並且在向前特徵選取法中挑選的特徵包含 eGFR、RBC、BUN 與 Creatinine(表三)。



圖六、向前特徵選取法的示意圖

表三、模型結果的評分比較

特徵：eGFR、RBC、BUN 與 Creatinine					
Models	**Accuracy (95% CI)**	**Precision (95%CI)**	**Sensitivity (95%CI)**	**Specificity (95%CI)**	**AUC (95%CI)**
Decision tree	79.1%(77.7%–80.5%)	82.7%(81.1%–84.2%)	92.4%(90.60%–94.1%)	33.7%(28.2%–39.2%)	0.760(0.74–0.779)
Random forest	79.5%(78.0%–80.9%)	83.5%(82.1%–84.9%)	91.5%(90.4%–92.7%)	37.9%(34.4%–41.5%)	0.801(0.785–0.816)
SVM	78.2%(76.7%–79.80%)	78.7%(77.0%–80.4%)	98.6%(97.7%–99.4%)	7.80%(3.40%–12.2%)	0.739(0.719–0.76)
XGBoost	79.7%(78.4%–81.10%)	83.4%(82.0%–84.8%)	92.2%(91.2%–93.3%)	36.6%(33.2%–40.0%)	0.807(0.792–0.823)
LightGBM	80.30%(78.8%–81.69%)	83.3%(82.0%–84.7%)	93.2%(92.10%–94.3%)	35.2%(31.9%–38.5%)	0.810(0.794–0.827)
Logistic regression	79.5%(78.10%–80.9%)	81.6%(80.2%–83.0%)	95.1%(94.2%–96.0%)	25.4%(22.2%–28.6%)	0.774(0.755–0.794)

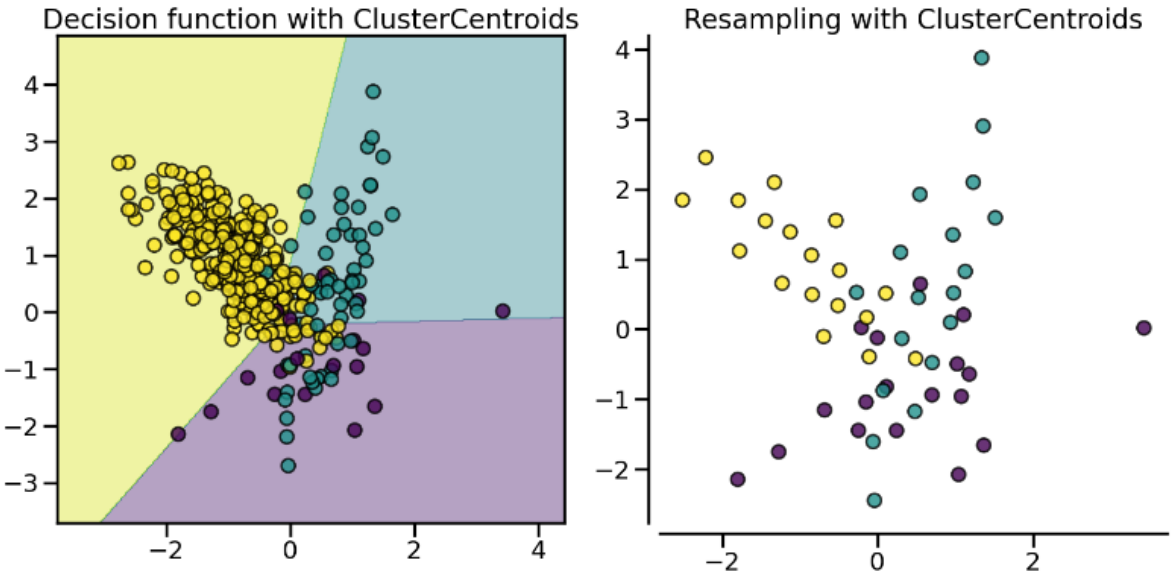
伍、討論

在研究中，從衛生福利部雙和醫院蒐集了三年臨床檢驗的大數據資料。針對取得的臨床數據含有缺失值的檔案(.csv)中的大腸直腸癌與大腸炎病患的資料進行資料後續處理。經移除掉了超過 50%缺失的項目，並且在剩餘的資料當中比較了不同補值方式對於資料的填補結果。發現在比較了補值後的資料分布與對於機器學模型訓練的結果後，使用 KNN 補值方法且 K=7 時擁有最佳的效能。

接下來使用了特徵排序的方式發現 eGFR 對於分辨大腸直腸癌和大腸炎病患的效能最好。在 Velciov 等人的研究中，也同樣的發現腎功能和大腸直腸癌具有高度的相關性，這與特徵排序的結果一致[24]。接續使用向前特徵選擇法，以 eGFR 項目為起始項目的資料，開始隨機的選擇特徵加入資料中，並且以驗證的結果為準，若模型的評分有增加則保留這個特徵。最終，發現 LightGBM 的效能最好，並且選取到的特徵為 eGFR、RBC、BUN 與 Creatinine。Burnett 等人對電子病歷結合機器學習模型進行大腸直腸癌風險預測的文獻回顧，使用模型包含 Decision tree、Random forest、Logistic regression、深度神經網絡(Deep neural network, DNN)、卷積神經網絡(Convolutional neural network, CNN)、人工神經網絡(Artificial neural network, ANN)、分類和回歸樹(Classification and regression tree, CART)，其精確度、靈敏度、特異性與接收者操作特徵曲線下面積為 3.0-98.0%、3.91-88.6%、82.73-95%與 0.686-0.93.7[25]。其中，Birks 等人使用組合模型(Decision tree 與 Random forest)及特徵選取(年齡、性別及各種血球數據)進行 25,430 人的大腸直腸癌風險預測，其靈敏度、特異性與接收者操作特徵曲線下面積為 3.91%、82.73%與 0.776 [26]。Cooper 等人使用 ANN 與糞便免疫化學試驗進行 1,810 人的大腸直腸癌風險預測，其靈敏度、特異性與接收者操作特徵曲線下面積為 35.15%、85.57%與 0.686 [27]。Wu 等人使用 Decision tree 及特徵選取(年齡、性別、身高、體重及 BMI)進行 225 人的大腸直腸癌風險預測，其靈敏度、特異性與接收者操作特徵曲線下面積為 82.5%、92.2%與 0.937[28]。另外，Chang 使用 LightGBM 及特徵選取(CEA 與 RBC)進行 7,747 人的大腸直腸癌風險預測，其準確度、精確度、靈敏度、特異性與接收者操作特

徵曲線下面積為 85.5%、86.3%、84.3%、86.6%與 0.919[29]。所以，機器學習模型對不同地區電子病歷、樣本數大小及特徵選取種類，對大腸直腸癌風險有不同的預測結果。

機器學習的實作中，相當重要的是關於機器學習的評分結果是否接近真實。舉例來說，訓練過程中，可能會因為資料帶有答案(如病患編號)或者樣本比例不均(病例過少)造成評分良好的假象。流行病學指出病例對照樣本數最適合的比例是 1:1~1:4，大腸直腸癌與大腸炎的樣本數比例是 1:2.23。研究發現分類結果異常可能帶來精確度低但是準確度高的問題，可能是大腸直腸癌樣本數過多，因此使用 Imblearn 這個套件來解決問題。此套件能夠將較多的樣本的數量在不改變資料分布的狀態下，減低樣本數，如圖七所示。所有評分結果都會得了 75%或 0.75 以上的評分，除了特異性的結果較差。推測可能是資料中所收取的項目不夠廣泛，或者缺乏與大腸直腸癌有直接關聯性的檢驗項目資料。然而，若考量到使用這個模式進行檢測，不僅能夠提升病患的接受篩檢率，而且這一個篩檢能夠提供 93.2%的靈敏度，仍然是一個值得發展的篩檢工具。



圖七、套件 Imblearn 中 Clustercentroid 的示意圖

陸、結論

本研究透過使用機器學習與醫院的大腸直腸癌與大腸炎的臨床資料結合，找出最能夠用以區分癌症的檢驗項目。機器學習在醫療領域的應用正在逐漸普及，並有望為醫療行業帶來許多正面的影響。隨著技術的進步和資料量的增加，機器學習在醫療領域的應用將會越來越廣泛。首先，機器學習技術可以幫助臨床快速且準確地診斷疾病，減少診斷錯誤的機率。例如，透過機器學習技術可以輔助醫生分析大量的影像資料、檢驗項目資料以及病歷資料，並快速篩選出患有某些疾病風險的病人，並提供給臨床進一步的診斷。此外，機器學習技術還可以幫助臨床量身訂製治療方案。例如，利用機器學習技術可以透過分析大量的醫療資料，找出治療某種疾病的最佳方案，並提供給臨床參考。最

後，機器學習技術還可以用於預測病人的預後。例如，透過分析病人的醫療資料和生活型態，可以預測病人患有某種疾病的機率，並提醒醫生及時進行治療。總而言之，機器學習在醫療領域的應用將會越來越廣泛，能夠為醫生提供更多的幫助，提高醫療品質。這類型的醫療稱之為精準醫療。精準醫療將成為未來醫療的重要方向，有利於精確診斷、選擇適當治療方案和監測病情。通過基因檢測、蛋白質檢測、代謝體檢測等技術，可以精確判斷患者的健康狀況，幫助醫生選擇適合患者的治療方案。

此外，精準醫療還將促進醫療的自動化和智慧化，通過人工智慧技術、機器學習和大數據分析，可以更快速和準確地診斷疾病，幫助醫生做出更精準的治療決策。隨著精準醫療技術的進步，醫療品質將得到更大的提升，患者將能得到更好的醫療服務，減少不必要的治療，提高治癒率和治療效果。

柒、參考文獻資料

1. Hirai, H.W., et al., *Systematic review with meta-analysis: faecal occult blood tests show lower colorectal cancer detection rates in the proximal colon in colonoscopy-verified diagnostic studies.* Aliment Pharmacol Ther, 2016. **43**(7): p. 755-64.
2. Luh, D.-L., *Relationships between colorectal cancer screening behavior and incentives in Taiwan-observational and intervention research (Final report in MOST 107-2410-H-040-011-SSS).* Department of Public Health, Chung Shan Medical University, 2019.
3. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.
4. Schaffer, C., *Selecting a classification method by cross-validation.* Machine Learning, 1993. **13**(1): p. 135-143.
5. Marom, N.D., L. Rokach, and A. Shmilovici. *Using the confusion matrix for improving ensemble classifiers*. in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*. 2010.
6. Rahmani, K., et al., *Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction.* medRxiv, 2022.
7. Singh, A., N. Thakur, and A. Sharma. *A review of supervised machine learning algorithms*. in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. 2016.
8. Usama, M., et al., *Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges.* IEEE Access, 2019. **7**: p. 65579-65615.
9. Somvanshi, M., et al. *A review of machine learning techniques using decision tree and support vector machine*. in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*. 2016.
10. Zeng, Y. and F. Cheng, *Medical and Health Data Classification Method Based on Machine Learning.* Journal of Healthcare Engineering, 2021. **2021**: p. 2722854.
11. Suthaharan, S., *Support Vector Machine*, in *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. 2016, Springer US: Boston, MA. p. 207-235.
12. *XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury.* Neurotrauma Reports, 2020. **1**(1): p. 8-16.
13. Ji, X., et al., *Prediction Model of Hypertension Complications Based on GBDT and LightGBM.* Journal of Physics: Conference Series, 2021. **1813**(1): p. 012008.
14. Greener, J.G., et al., *A guide to machine learning for biologists.* Nature Reviews Molecular Cell Biology, 2022. **23**(1): p. 40-55.
15. Zien, A., et al. *The Feature Importance Ranking Measure*. 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.
16. Huynh-Thu, V.A., et al., *Statistical interpretation of machine learning-based feature importance scores for biomarker discovery.* Bioinformatics, 2012. **28**(13): p. 1766-1774.
17. Ledezma, C.A., et al., *A modeling and machine learning approach to ECG feature engineering for the detection of ischemia using pseudo-ECG.* PLOS ONE, 2019. **14**(8): p. e0220294.
18. Ebiaredoh-Mienye, S.A., et al., *A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease.* Bioengineering (Basel), 2022. **9**(8).

19. Behura, A., *The Cluster Analysis and Feature Selection: Perspective of Machine Learning and Image Processing*, in *Data Analytics in Bioinformatics*. 2021. p. 249-280.
20. Saberi-Karimian, M., et al., *Potential value and impact of data mining and machine learning in clinical diagnostics.* Critical Reviews in Clinical Laboratory Sciences, 2021. **58**(4): p. 275-296.
21. Domínguez-Olmedo, J.L., et al., *Machine Learning Applied to Clinical Laboratory Data in Spain for COVID-19 Outcome Prediction: Model Development and Validation.* J Med Internet Res, 2021. **23**(4): p. e26211.
22. Bai, Q., et al., *Machine learning to predict end stage kidney disease in chronic kidney disease.* Scientific Reports, 2022. **12**(1): p. 8377.
23. Gould, M.K., et al., *Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data.* Am J Respir Crit Care Med, 2021. **204**(4): p. 445-453.
24. Velciov, S., et al., *Aspects of renal function in patients with colorectal cancer in a gastroenterology clinic of a county hospital in Western Romania.* Rom J Intern Med, 2013. **51**(3-4): p. 164-71.
25. Burnett, B., et al., *Machine Learning in Colorectal Cancer Risk Prediction from Routinely Collected Data: A Review.* Diagnostics (Basel), 2023. **13**(2).
26. Birks, J., et al., *Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records.* Cancer Med, 2017. **6**(10): p. 2453-2460.
27. Cooper, J.A., et al., *Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model.* Br J Cancer, 2018. **118**(2): p. 285-293.
28. Wu, H.C., et al., *Developing screening services for colorectal cancer on Android smartphones.* Telemed J E Health, 2014. **20**(8): p. 687-95.
29. Chang, Y.-T., *Using Clinical Laboratory Data of Colorectal Cancer By Importing Machine Learning Algorithm To Build Early Cancer Prediction Model (Master Thesis).* Taipei Medical University, College of Medical Science and Technology, Professional Master Program in Medical Laboratory Science and Biotechnology, 2023.

【評語】052512

1. 本作品以三年臨床檢驗的大數據資料，針對臨床數據含有缺失值的大腸直腸癌與大腸炎病患的資料進行資料後續處理。實驗過程比較了不同補值方式對於資料的填補效果，實驗顯示 KNN 補值方法且 K=7 時擁有最佳的效能。整體研究完整度高。

2. 本作品是很典型的機器學習應用，資料來源與數量雖是重要的因素，但是機器學習模型的選擇與實驗也甚為重要。且本作品隸屬「電腦與資訊學科」組，建議可多增加資訊技術創新、研究等的分析探討，而非僅以應用成效分析為主。

作品海報

以大腸直腸癌預測為例進行缺失值處理方式的探討與實驗

壹、研究動機

前人已有許多利用機器學習和醫院的檢驗項目資料來發展應用並達到針對疾病的早期預測的研究，代表著這類型研究的可行性很高。因此本次的研究目的是要以常規檢驗項目做出能夠預測大腸直腸癌的機器學習模型。

貳、研究目的

主要目的: 健康檢查的血液、尿液檢測項目做出機器學習預測模型

1. 哪一種補值得方法最合適?
2. 哪一些檢測項目最適合用於預測大腸直腸癌?
3. 是否可以透過處理缺失值、特徵選擇加強機器學習的表現?

參、研究過程及方法

【實驗一】資料處理與缺失值處理

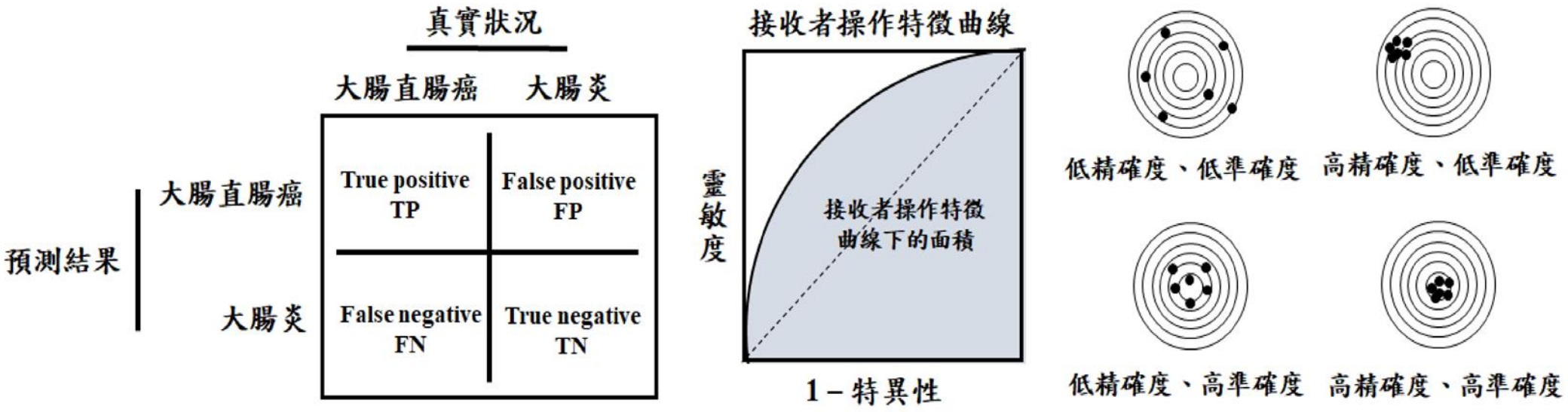
從臨床實驗室檢驗項目中，搜集了共56個檢驗項目結果。若檢驗項目的缺失比例達整體病人的50%，則將該檢驗項目移除。處理過後，剩餘22個檢驗項目，其缺失比例為0 到44.1% (表一)。缺失的資料的處理方式以補值、刪除、以及KNN插補法補值。補值是以平均值(連續性變項)、中位數(類別變項)填補。刪除是以若病人有任一缺失，則將該病人移除。KNN插補法進行補值則是利用無缺失的鄰近值，計算缺失的數值。先以學生t檢定的方式比較原始資料與填補資料是否達到統計上的顯著差異，再由邏輯式回歸(logistic regression)訓練與驗證之準確度的標準差結果進行比較。

【實驗二】特徵選取

使用WEKA(v. 3.8.3)進行特徵排序。

【實驗三】模型訓練與驗證

本研究以向前特徵選取法來進行特徵選取並使用 10倍交叉驗證將實驗數據分成訓練資料集與驗證資料集。使用scikit learn 0.21.3 框架中使用決策樹(decision tree)、隨機森林(random forest)、支援向量機(support vector machine, SVM)、極限梯度提升(eXtreme gradient boosting, XGBoost)、輕量化梯度提升機(light gradient boosting machine, Light GBM)與邏輯式回歸，每一個模型都經過測試以找到最適合的參數條件使用。參數測試決策樹中測試了kernel (gini, entropy)與depth (1 ~ 10)，而在隨機森林中測試了kernel (gini, entropy)與tree number (100 ~ 500)。在SVM中測試了gamma值(1e-6 ~ 1e-10)與C值(1e5 ~ 1e7)。在XGBoost中測試了eta值(0.01 ~ 0.2)與depth (1 ~ 10)，然後在LightGBM中測試了leaves (50 ~ 400)與depth (1 ~ 10)。研究經反覆測試，以較佳的準確度(Accuracy)作為進步依據進行模型的調整，並且最終將預測正確與錯誤的值整合，以準確度、精確度(Precision)、靈敏度(Sensitivity)、特異性(Specificity)與接收者操作特徵曲線下的面積(AUC)來呈現模型評估結果(圖一)。



$$\text{準確度} = \text{人群(大腸直腸癌+大腸炎)中，準確預測是大腸直腸癌的比率} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{靈敏度} = \text{罹患大腸直腸癌患者，預測是大腸直腸癌的比率} = \frac{TP}{TP+FN}$$

$$\text{特異性} = \text{罹患大腸炎患者，預測是大腸炎的比率} = \frac{TN}{FP+TN}$$

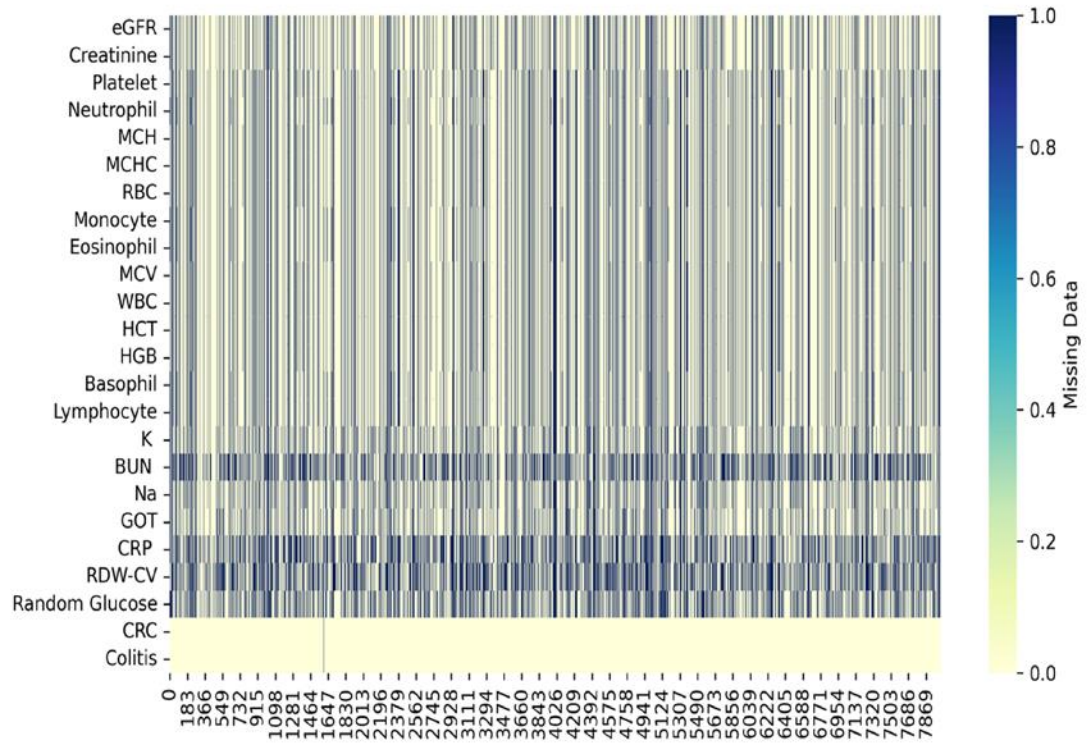
$$\text{精確度} = \text{預測是大腸直腸癌患者，診斷試驗結果是大腸直腸癌的比率又稱陽性預測值} = \frac{TP}{TP+FP}$$

圖一、混淆矩陣與接收者操作特徵曲線

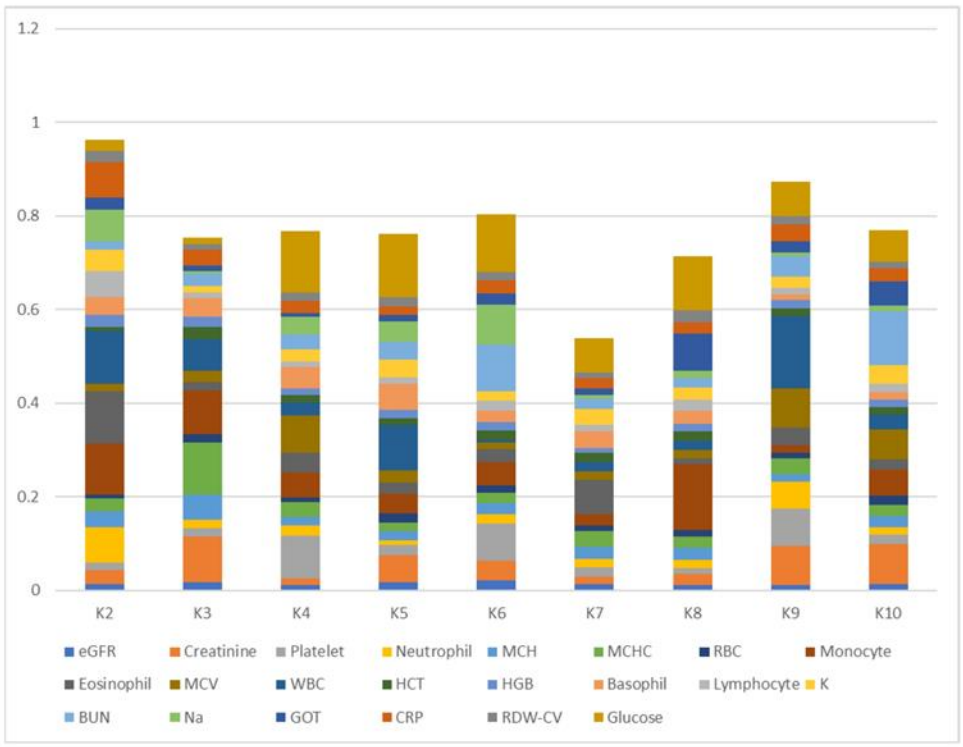
肆、研究結果

【實驗一】資料前處理與缺失值處理

經由台北醫學大學附設醫院(IRB編號: N202201049)蒐集資料，總共取得24863筆資料。整理後發現共有8047筆病人的資訊，圖二缺失值的視覺化呈現。本次研究只取用檢驗項目的資訊，>50%的病人缺少的檢驗項目被移除，最終剩下6792筆病人的資料，其中大腸炎病人為2105 位及大腸直腸癌病人為4687位。表一是原始資料與補值後資料的比較，圖三是以邏輯式回歸模型進行訓練時累進的標準差結果。



圖二、缺失值的視覺化圖



圖三、以邏輯式回歸模型進行訓練時累進的標準差結果

表一、原始資料與補值後資料的比較

	Origin CRC	Delete CRC	Mean CRC	K2	K3	K4	K5	K6	K7	K8
eGFR	**135.63±157.93**	117.08±146.22 **	135.63±139.36	135.84±140.77	135.67±140.41	136±140.93	136.13±140.76	136.1±140.48	136.09±140.36	136.02±140.23
Creatinine	**1±1.14**	1.3±1.66 **	1±1.01	0.99±1.01	0.99±1.01	0.99±1.01	0.99±1.01	0.99±1.01	0.99±1.01	0.99±1.01
Platelet	**241.08±90.96**	235.5±107.25 *	241.11±78.36	242.1±80.04	242.29±79.59	242.11±79.27	242.14±79.08	242.25±79.05	242.22±78.98	242.19±78.93
Neutrophil	**69.48±15.61**	69.39±15.91	69.48±13.43	69.34±13.84	69.32±13.71	69.26±13.68	69.27±13.62	69.27±13.6	69.28±13.57	69.28±13.56
MCH	**29.47±3.36**	29.75±3.46 *	29.47±2.9	29.47±2.94	29.47±2.93	29.48±2.93	29.48±2.92	29.48±2.92	29.48±2.91	29.48±2.91
MCHC	**34.01±1.02**	33.94±1.07 *	34.01±0.88	34.03±0.9	34.03±0.89	34.03±0.89	34.03±0.89	34.03±0.89	34.03±0.89	34.03±0.88
RBC	**4.48±0.8**	4.18±0.89 **	4.48±0.69	4.5±0.7	4.5±0.7	4.5±0.69	4.5±0.69	4.5±0.69	4.5±0.69	4.5±0.69
Monocyte	**7.76±3.65**	7.83±4.04	7.76±3.14	7.76±3.2	7.77±3.18	7.77±3.17	7.76±3.17	7.76±3.16	7.76±3.16	7.76±3.16
Eosinophil	**1.69±2.11**	1.9±2.39 *	1.69±1.81	1.69±1.85	1.69±1.84	1.69±1.83	1.69±1.83	1.69±1.83	1.69±1.82	1.69±1.82
MCV	**86.53±8.69**	87.52±8.94 **	86.53±7.49	86.5±7.61	86.5±7.57	86.52±7.56	86.52±7.55	86.51±7.54	86.51±7.53	86.52±7.53
WBC	**9.51±4.9**	9.53±5.93	9.51±4.22	9.49±4.29	9.49±4.26	9.49±4.26	9.49±4.25	9.5±4.24	9.5±4.24	9.5±4.24
HCT	**38.5±6.26**	36.23±7.03 **	38.49±5.4	38.65±5.49	38.66±5.47	38.67±5.47	38.67±5.46	38.67±5.46	38.67±5.45	38.66±5.45
HGB	**13.11±2.22**	12.31±2.46 **	13.11±1.91	13.16±1.95	13.17±1.94	13.17±1.94	13.17±1.94	13.17±1.94	13.17±1.93	13.17±1.93
Basophil	**0.48±0.39**	0.49±0.42	0.48±0.33	0.48±0.34	0.48±0.34	0.48±0.34	0.48±0.34	0.48±0.34	0.48±0.34	0.48±0.34
Lymphocyte	**20.27±13.1**	19.86±13.32	20.27±11.27	20.42±11.64	20.44±11.53	20.5±11.49	20.5±11.45	20.5±11.42	20.49±11.4	20.49±11.39
K	**3.82±0.54**	3.86±0.64 *	3.82±0.45	3.82±0.47	3.82±0.46	3.82±0.46	3.82±0.46	3.82±0.46	3.82±0.46	3.82±0.46
BUN	**18.69±19.26**	21.97±23.45 **	18.69±14.2	17.07±14.6 **	17.07±14.58 **	17.07±14.56 **	17.06±14.54 **	17.06±14.54 **	17.06±14.55 **	17.06±14.55 **
Na	**137.75±3.77**	137.65±4.67	137.75±3.2	137.82±3.26	137.83±3.24	137.82±3.23	137.82±3.23	137.82±3.22	137.82±3.22	137.81±3.22
GOT	**36.54±215.37**	46.6±356.7	36.54±184.98	35.98±185.22	35.97±185.16	35.91±185.06	35.89±185.04	35.87±185.03	35.86±185.02	35.86±185.02
CRP	**3.66±5.69**	4.41±6.56 **	3.66±4.22	3.29±4.49 *	3.31±4.42 *	3.3±4.37 *	3.29±4.35 *	3.29±4.34 *	3.29±4.32 *	3.3±4.32 *
RDW-CV	**14.61±2.55**	15.18±2.91 **	14.61±1.86	14.4±2.03 **	14.4±2 **	14.4±1.98 **	14.4±1.97 **	14.4±1.96 **	14.4±1.96 **	14.41±1.96 **
Random Glucose	**130.05±60.76**	142.67±74.46 **	130.05±48.04	125.63±50.1 **	125.79±49.69 **	125.98±49.45 **	126.26±49.28 *	126.45±49.15 *	126.62±49.06 *	126.73±48.98 *

*** $p < 0.05$, ** $p < 0.001$**

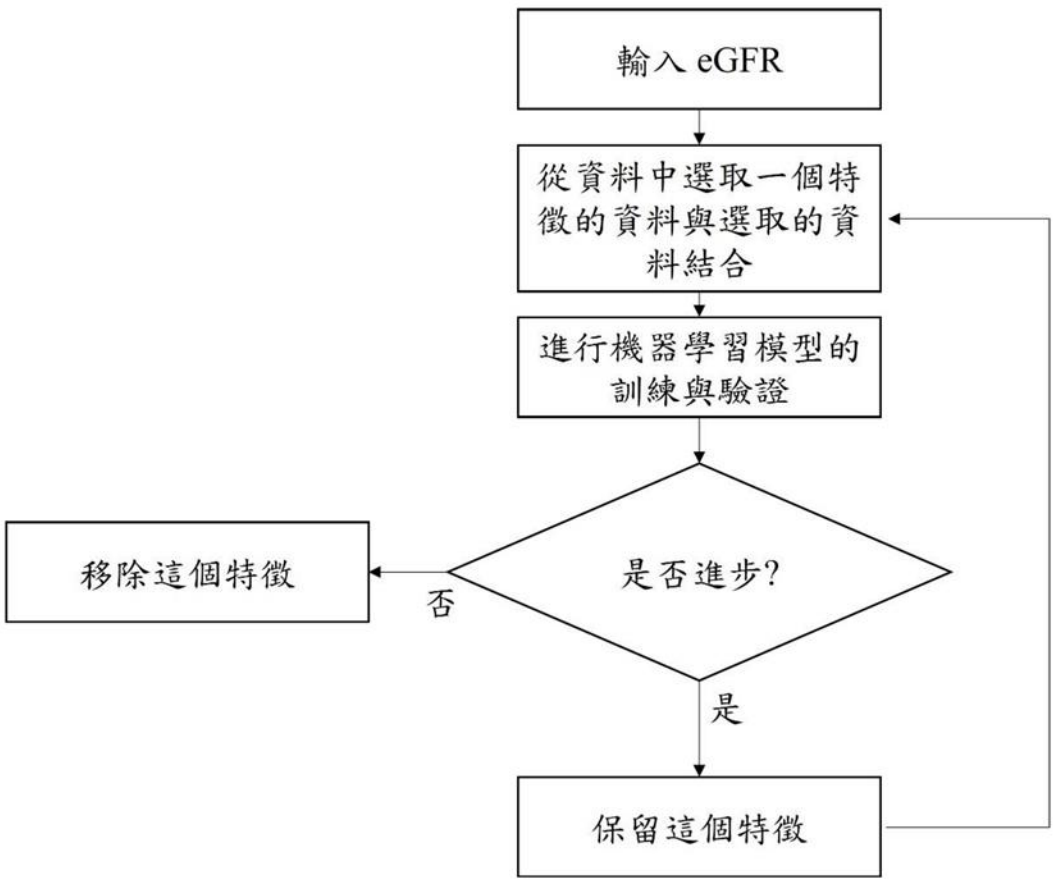
【實驗二】特徵選取

表二、使用WEKA進行特徵排序

Ranking Score	Attributes
0.0894	**eGFR**
0.0832	RBC
0.0821	BUN
0.0748	RDW-CV
0.0643	HGB
0.0613	HCT
0.0572	Creatinine
0.0443	MCV
0.0435	MCH
0.0431	Platelet
0.042	WBC
0.0353	CRP
0.0347	Na
0.0313	K
0.028	Eosinophil
0.0277	Basophil
0.0269	Random Glucose
0.0222	MCHC
0.0221	GOT
0.0209	Lymphocyte
0.0195	Monocyte
0.0162	Neutrophil

【實驗三】模型訓練與驗證

使用向前特徵選取法(圖四)選取特徵，完成模型結果的評分表較(表三)



圖四、向前特徵選取法的示意圖

表三、模型結果的評分比較

特徵：eGFR、RBC、BUN 與 Creatinine					
Models	**Accuracy (95% CI)**	**Precision (95%CI)**	**Sensitivity (95%CI)**	**Specificity (95%CI)**	**AUC (95%CI)**
Decision tree	79.1%(77.7%–80.5%)	82.7%(81.1%–84.2%)	92.4%(90.60%–94.1%)	33.7%(28.2%–39.2%)	0.760(0.74–0.779)
Random forest	79.5%(78.0%–80.9%)	83.5%(82.1%–84.9%)	91.5%(90.4%–92.7%)	37.9%(34.4%–41.5%)	0.801(0.785–0.816)
SVM	78.2%(76.7%–79.80%)	78.7%(77.0%–80.4%)	98.6%(97.7%–99.4%)	7.80%(3.40%–12.2%)	0.739(0.719–0.76)
XGBoost	79.7%(78.4%–81.10%)	83.4%(82.0%–84.8%)	92.2%(91.2%–93.3%)	36.6%(33.2%–40.0%)	0.807(0.792–0.823)
LightGBM	80.30%(78.8%–81.69%)	83.3%(82.0%–84.7%)	93.2%(92.10%–94.3%)	35.2%(31.9%–38.5%)	0.810(0.794–0.827)
Logistic regression	79.5%(78.10%–80.9%)	81.6%(80.2%–83.0%)	95.1%(94.2%–96.0%)	25.4%(22.2%–28.6%)	0.774(0.755–0.794)

伍、討論

【實驗一】資料前處理與缺失值處理

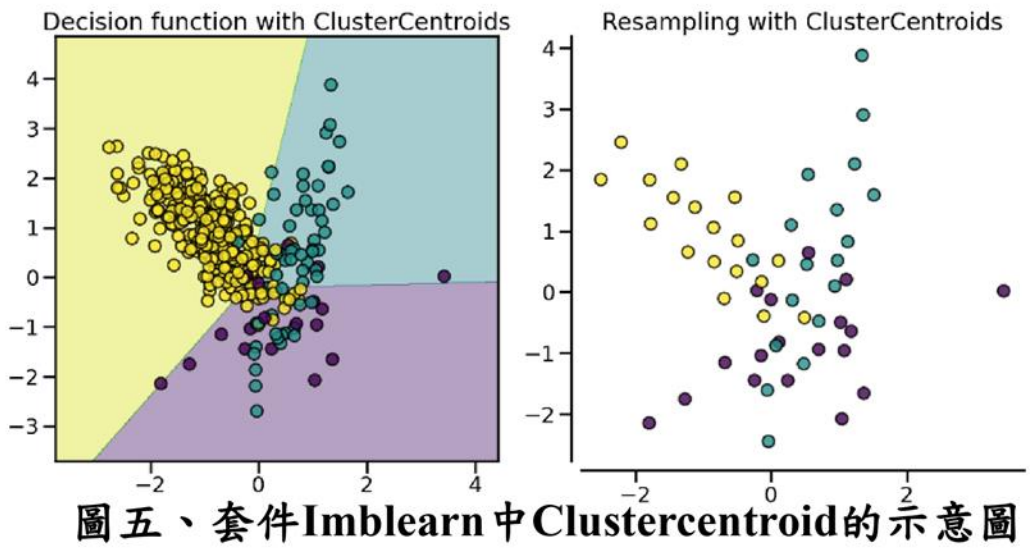
在研究中從台北醫學大學附設醫院蒐集回顧了三年的實驗診斷資料。針對資料中的大腸直腸癌病患與大腸炎病患的資料進行資料前處理。移除掉了超過50%缺失的項目，並且在剩餘的資料當中比較了不同補值方式對於資料的填補結果。發現在比較了補值後的資料分布與對於機器學模型訓練的結果後，使用KNN補值方法且K=7時擁有最佳的效果。

【實驗二】特徵選取

使用特徵排序的方式發現eGFR對分辨大腸直腸癌和大腸炎的效果最好。在Velciov等人研究中也發現腎功能和大腸直腸癌具有高度的相關性，這與特徵排序的結果一致。

【實驗三】模型訓練與驗證

接續使用向前特徵選擇法，以eGFR項目為起始項目的資料，開始隨機選擇特徵加入資料中，以驗證的結果為準，若模型的評分有增加則保留此特徵。最終發現LightGBM的效果最好，並且選取的特徵為eGFR、RBC、BUN與Creatinine。機器學習的實作中，相當重要的是機器學習的評分結果是否接近真實。舉例來說，訓練過程中可能會因為資料帶有答案(如病患編號)或者樣本比例不均(陽性過少)造成評分良好的假象。實作當中有考量進樣本比例不均可能帶來的問題，因此使用Imblearn這一個套件來解決這個問題(圖五)。這個套件能夠將較多的樣本的數量在不改變資料分布的狀態下，減低樣本數。所有評分結果都會得了75%或0.75以上的評分，除了特異性的結果較差。推測可能是資料中所收取的項目不夠廣泛，或者缺乏與大腸直腸癌有直接關聯性的檢驗項目資料。然而，若考量到使用這個模式進行檢測，不僅能夠提升病患的接受篩檢率，這個篩檢能夠提供93.2%靈敏度，仍然是一個值得發展的篩檢工具。



圖五、套件Imblearn中Clustercentroid的示意圖

陸、結論

本研究使用機器學習與醫院的大腸直腸癌與大腸炎的臨床資料結合，找出最能夠區分癌症的檢驗項目。機器學習在醫療領域的應用正在逐漸普及，並為醫療行業帶來許多正面的影響。隨著技術的進步和資料量的增加，機器學習在醫療領域的應用將會越來越廣泛。首先，機器學習可以幫助臨床快速且準確地診斷疾病，減少診斷錯誤的機率。例如，透過機器學習可以輔助醫生分析大量的影像資料、檢驗項目資料以及病歷資料，並快速篩選患有某些疾病風險的病人，並提供給臨床進一步的診斷。此外，機器學習還可以幫助臨床量身訂製治療方案。例如，利用機器學習可以透過分析大量的醫療資料，找出治療某種疾病的最佳方案，並提供給臨床參考。最後，機器學習還可以用於預測病人的預後。通過基因檢測、蛋白質檢測、代謝體檢測等技術，可以精確判斷患者的健康狀況，幫助醫生選擇適合患者的治療方案。隨著精準醫療技術的進步，患者將能得到更好的醫療服務，減少不必要的治療，提高治癒率和治療效果。

柒、參考資料

1. LeCun, Y., Y. Bengio, and G. Hinton, Deep learning. Nature, 2015. 521(7553): p. 436-444.
2. Schaffer, C., Selecting a classification method by cross-validation. Machine Learning, 1993. 13(1): p. 135-143.
3. Marom, N.D., L. Rokach, and A. Shmilovici. Using the confusion matrix for improving ensemble classifiers. In 2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel. 2010.
4. Rahmani, K., et al., Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. medRxiv, 2022.
5. Singh, A., N. Thakur, and A. Sharma. A review of supervised machine learning algorithms. in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). 2016.
6. Usama, M., et al., Unsupervised Machine Learning for Networking: Techniques, Applications, and Research Challenges. IEEE Access, 2019. 7: p. 65579-65615.
7. Somvanshi, M., et al. A review of machine learning techniques using decision tree and support vector machine. in 2016 International Conference on Computing Communication Control and Automation (ICCUBEA). 2016.
8. Zeng, Y., and F. Cheng, Medical and Health Data Classification Method Based on Machine Learning. Journal of Healthcare Engineering, 2021. 2021: p. 2722854.
9. Suthaharan, S., Support Vector Machine, in Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. 2016, Springer US: Boston, MA. p. 207-235.
10. XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury. Neurotrauma Reports, 2020. 1(1): p. 8-16.
11. Ji, X., et al., Prediction Model of Hypertension Complications Based on GBDT and LightGBM. Journal of Physics: Conference Series, 2021. 1813(1): p. 012008.
12. Zien, A., et al. The Feature Importance Ranking Measure. 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.
13. Huynh-Thu, V.A., et al., Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics, 2012. 28(13): p. 1766-1774.
14. Velciov, S., et al., Aspects of renal function in patients with colorectal cancer in a gastroenterology clinic of a county hospital in Western Romania. Rom J Intern Med, 2013. 51(3-4): p. 164-71.