運用機器學習研究前列腺癌診斷模型

**亞洲大學-生物資訊與醫學工程學系-3A-107025026-邱筱妍**

**大綱**

前列腺癌是僅有男性有的癌症。近幾年以來發生率及死亡率都有明顯上升，這種現象在高度開發國家較為明顯，有可能是因為廣泛的前列腺癌篩檢及年齡高齡老化有關。本研究將在Kaggle中獲得的大量前列腺癌臨床數據，運用python做機器學習，對機器進行學習訓練，並找出可能與前列腺癌有關的特徵，希望能讓前列腺癌的醫療診斷更加精確快速。

1. **前言**

前列腺癌（Prostate Cancer）是男性癌症死亡的第二大元兇。許多男性很容易忽略。在臺灣，根據統計從民國69年，每年每十萬人就至少有100人罹患前列腺癌。到民國104年，達到將近5,000人，整整成長了50倍。而且自民國104年以來，前列腺癌的發病率一直位居所有男性惡性腫瘤的前5名。依據增加速度來看，很有可能在幾十年內成為男性前三大癌症。早期診斷及治療對於提高前列腺癌患者生存率的關鍵，但在前列腺癌的診斷和治療過程中一直都存在著過度診斷和治療的問題。有可能造成前列腺癌發生的原因有年齡大於50歲、基因變異、家族史、肥胖、攝護腺炎等等，甚至有學者發現位居不同地區發生罹癌機率也不一樣，有可能跟生活習慣息息相關。

機器學習是透過資料訓練機器。機器從資料中得到複雜的函數來學習並創造一組規則，並利用它來做預測。機器學習共有三種學習方式。1. 監督式學習：電腦從已經處理過的標準化資訊中分析模式後做出預測的學習方式。2. 非監督式學習：訓練資料不需要事先處理，讓機器自己學習處理資料，依照關聯性去歸類，再找出規則，不會有正確或不正確的判別。3. 強化式學習：不需要給機器任何的資料，讓機器直接透過環境的正向、負向回饋，從互動中去學習，這是最接近人類原本的學習方式。本研究將運用機器學習的方式，來研究前列腺癌診斷模型及特徵。

1. **方法**

首先，在Kaggle尋找並下載前列腺癌的臨床統計資料，資料內容包含診斷結果。本研究使用Anaconda中的Jupyter打Python3。Python語言強調程式碼的可讀性及簡潔的語法，它優雅的語法和動態類型，使它成為大多數平台上及許多領域程式編寫和應用程序開發的理想程式語言。先在Python中載入下載的資料，並列印出資料資訊，查看裡面有甚麼資訊。刪除一些不必要的資料。查看總共有幾筆資料、幾筆惡性、幾筆良性。在運用Seaborn可視化庫，繪製熱力圖，顯示變量的相關係數矩陣。再運用Seaborn繪製散佈圖矩陣，可以得知所有變量為正相關還是負相關，還有變量關係強度。 再運用Seaborn繪製盒狀圖，又稱四分位圖，來顯示資料在各變量中的分散程度。選出兩個最好的前列腺癌特徵。

將數據集劃分為訓練集及測試集，是為了在數據分析的過程中，保證模型在實際系統中能夠起到預期的作用。再來運用sklearn中的SVM演算法，1. SVC(Support Vector Classification): 支援向量分類，基於libsvm實現的，資料擬合時間複雜度是資料樣本的二次方。2. NuSVC（Nu-Support Vector Classification）: 支援向量分類，和SVC類似，不同的是通過一個引數空值支援向量的個數。3. LinearSVC(線性SVC):它具有更大的靈活性在選擇處罰和損失函式時，而且可以適應更大的資料集，他支援密集和稀疏的輸入是通過一對一的方式解決的。最後對分類模型的準確度做測試。

1. **結果**

圖一先用import as導入變數。

並用data = pd.read\_csv('下載檔案路徑');導入檔案，並印出檔案資訊。(圖二)

一張含有 文字 的圖片

自動產生的描述

(圖一)

一張含有 桌 的圖片

自動產生的描述

(圖二)

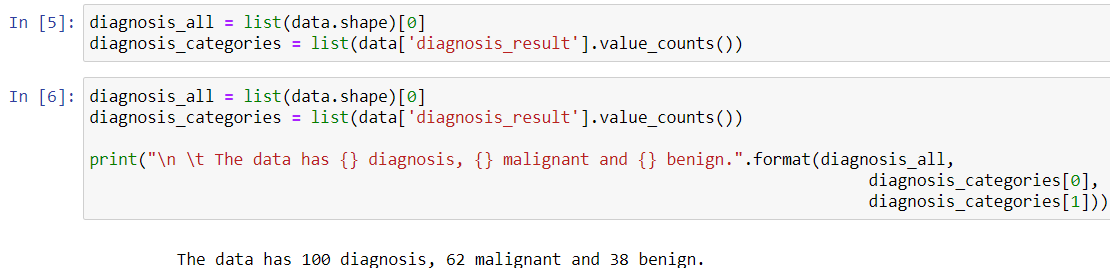
由圖三可得知檔案內共有100份資料，及10項主要指標，包含編號、診斷結果、半徑、質地、周長、區域、光滑度、緊密度、對稱性及分形維度。用data.drop()將沒有用處的資料刪除。再來圖三查看檔案中共有幾筆資料、幾筆良性、幾筆惡性。

一張含有 文字 的圖片

自動產生的描述

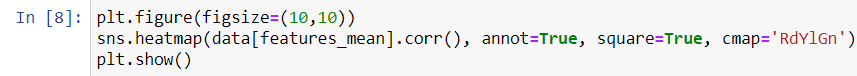
(圖三)

圖四可知檔案內共有100筆資料，其中有62筆惡性及38筆良性。



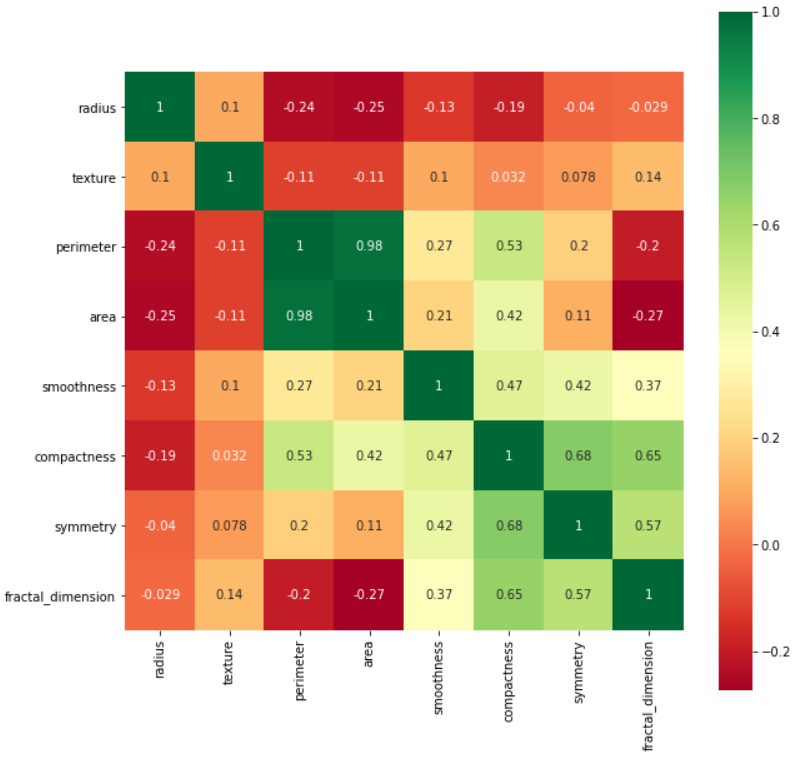
(圖四)

圖五為使用Seaborn可視化庫，繪製熱力圖，顯示變量的相關係數矩陣。設定圖中圖像區域大小為10英寸，輸出熱力圖數值大小，變成正方形，顏色為RdYlGn。



(圖五)

圖六中可發現直徑及區域的相關係數較大。



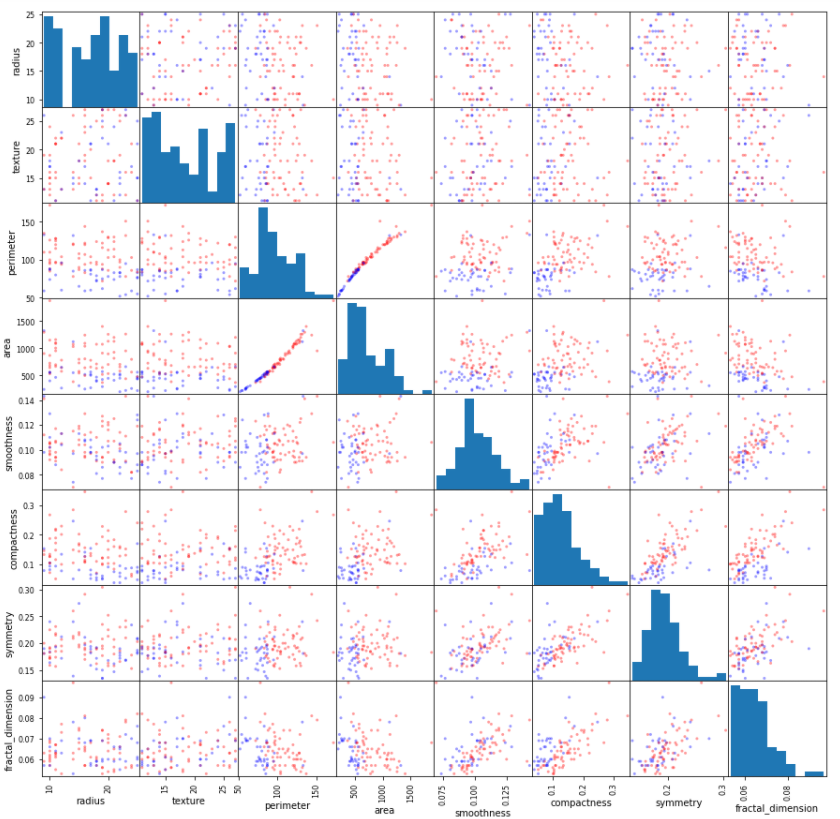
(圖六)

圖七為用Seaborn可視化庫，繪製散佈圖矩陣。在探索資料時，能看到不同變數間的散佈圖，也能直接看到不同變數間的關係。設診斷為惡性的為紅色、良性為藍色，顏色為colors，圖像透明度為0.4，圖像區域大小為15英寸。圖八為結果顯示，其中還是直徑及區域的相關性較大。

一張含有 文字 的圖片

自動產生的描述

(圖七)



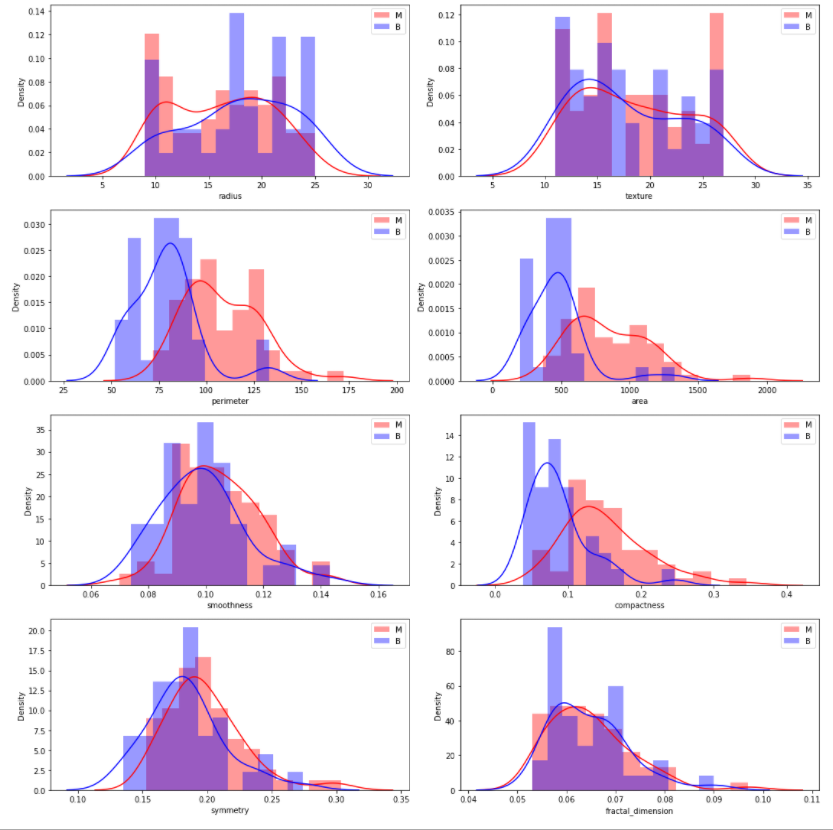
(圖八)

圖九為用Seaborn可視化庫，繪製分布密度圖。在探索資料時，可以更好地去評估分布的組成。設圖像區域大小為15英寸，顏色診斷惡性的為紅色、良性為藍色，長條圖邊界設為bins。圖十為結果顯示，其中直徑、區域及緊密度的惡性及良性重疊性較低，有可能可以成為判斷特徵。

一張含有 文字 的圖片

自動產生的描述

(圖九)



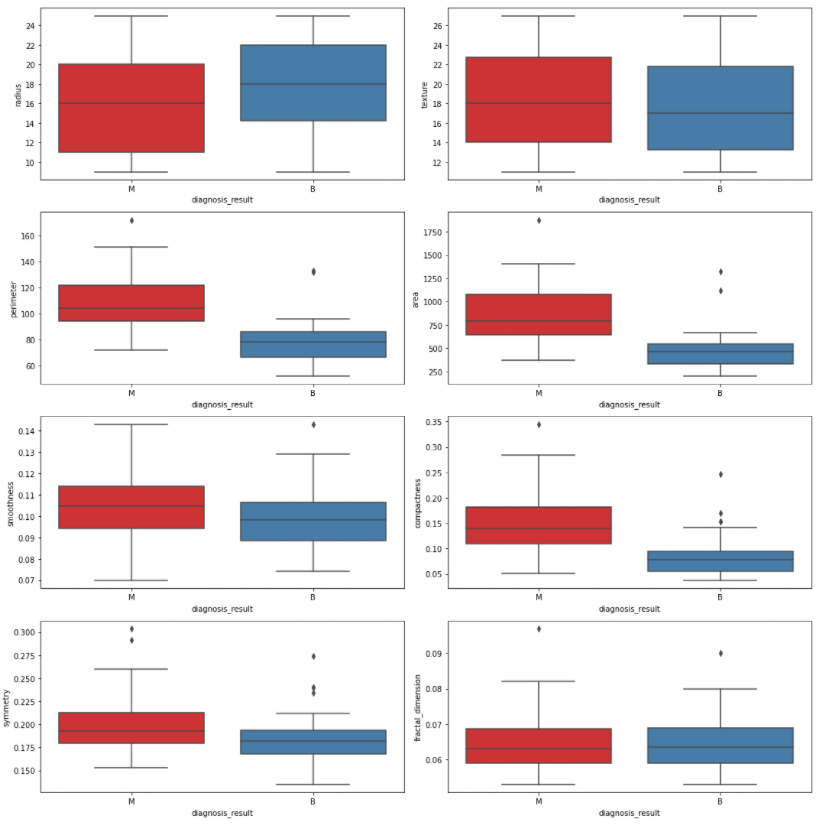
(圖十)

圖十一為用Seaborn可視化庫，繪製箱型圖。在探索資料時，可以顯示一組資料的最大值、最小值、上下四分位數、中位數、離群值及分散情況。設圖像區域大小為15英寸，x軸為診斷結果，y軸為特徵。圖十二為結果顯示，其中直徑、區域及緊密度的惡性及良性相似性較低，有可能可以成為判斷特徵。

一張含有 文字 的圖片

自動產生的描述

(圖十一)



(圖十二)

最後選定直徑及區域為選定特徵。(圖十三)

一張含有 文字 的圖片

自動產生的描述

(圖十三)

圖十四載入需要的資料庫，用accuracy\_score函數計算準確率，map()函數將方程式應用於可迭代的每一個元素，結果以列表的形式返回。將資料集分為訓練集及測試集。

一張含有 文字 的圖片

自動產生的描述

(圖十四)

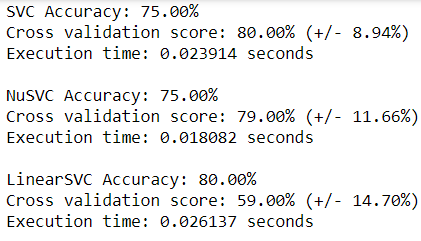
圖十五，計算SVC、NuSVC及LinearSVC所需要花費的時間，將資料標準化後做精準度預測。

一張含有 文字 的圖片

自動產生的描述

(圖十五)

圖十六為預測結果，可發現NuSVC花費時間較短，LinearSVC準確性較高，則SVC的交叉驗證性較高。交叉驗證的目的，是用未用來給模型作訓練的新數據，測試模型的性能，以便減少像過擬合和選擇偏差等問題，並給出模型如何在一個獨立的數據集上通用化。



(圖十六)

圖十七為最後對模型準確率做測試，這組模型準確率為76%。

一張含有 桌 的圖片

自動產生的描述

(圖十七)

1. **結論**

研究結果顯示前列腺癌的惡性及良性或許可以用腫瘤的直徑或是涵蓋區域去做判斷，所有準確率測驗雖然都有超過70%，但還是不入預期的高，模型的準確率也只有76%。臨床統計至今為止，前列腺癌過度診斷的比例還是有30%~50%這麼高，而且篩檢效果尚不明確。前列腺特異抗原檢測雖然可以增加癌症檢測率，但無法降低死亡率。由於大多數確診的前列腺癌的患者幾乎都沒有任何症狀，所以一些罹癌症患者並沒有進行深度的篩檢，可能導致他們一生都不會發現自己罹患了前列腺癌。由這次研究可以了解精準醫療的發展對於所有病症是一件非常有意義且重要的事，尤其是癌症，對於前列腺癌的臨床研究值得往後繼續探討。

機器學習可能只能做到讓機器從已經知道的資訊去做判斷及預測，而深度學習能讓機器從已經知道的資訊或未知的資訊中去產生並判斷連人類都沒找到的關鍵特徵，進而得到更加精準的預測，這兩者機器學習是相輔相成的，它們最終的目的都是希望機器能達到真正有用且精準的人工智慧。如何用現有的網路資源及專業的程式能力解決醫學上的問題，這就是生物資訊要去探討的地方。這堂課我學習到很多課堂上的知識，也學習到很多學資訊的應該要有的態度，多分享自己的想法，多與人交流，很謝謝王昭能教授耐心且無私的指導。

1. **參考文獻**
2. 王珏;歸納機器學習[A];2001年中國智能自動化會議論文集（上冊）[C];2001年
3. Cortes, C. , Vapnik, V. 1995. Support-vector networks. Machine Learning, 20: 273–297.
4. D'Amico AV, Whittington R, Malkowicz SB et al. . Pretreatment nomogram for prostate-specific antigen recurrence after radical prostatectomy or external-beam radiation therapy for clinically localized prostate cancer. J Clin Oncol 1999;17:168–172.
5. Jemal A, Siegel R, Ward E et al. . Cancer statistics, 2009. CA Cancer J Clin 2009;59:225–249.