利用初階健檢資料透過機器學習方法預測患者血管是否狹窄

生物機器學習 期末書面報告-第 16 組 許瑜朔

目錄

一、摘要 Abstract

背景：

心血管疾病指的是關於心臟或血管的疾病，其中冠狀動脈心臟病（簡稱冠心病）是最常見的一種心臟病，其負責供應心臟血液與養分，然而若血管發生狹窄或堵塞，血液與養分不易通過，就容易造成心臟缺氧而致死。當血管壁上出現過量鈣的現象時會造成血管動脈硬化及血管狹窄，容易造成心肌梗塞和腦中風等危害。

「十年心血管疾病風險評估」能針對尚未發生因粥狀動脈硬化造成心血管疾病之成年人，利用年齡、抽菸習慣、是否有糖尿病、血壓、血脂肪（總膽固醇、低密度脂蛋白膽固醇、高密度脂蛋白膽固醇）等六項指標，建立心血管風險預測模型，預測出未來十年可能罹患缺血性心臟病的機率，是目前在臨床上最常使用的評估方法。

方法：

利用初階健康檢查的資料及加入佛萊明漢風險評分(FRS)等欄位後的健檢資料，透過多樣機器學習方法（WEKA - SMO、Naïve Bayes、Random Forest、Logistic Regression、LibSVM、XGBoost 及 IBCGA+SVM）預測患者血管是否狹窄，比較各實驗在不同方法下的模型結果並進行討論。實驗一為原始數據（無 FRS 等欄位）的資料，實驗二是原始數據加上 FRS 等欄位的資料，實驗三則是將原始數據無缺值的資料放入 train 並將有缺值經 KNN 補值後放入 test，兩者都加上 FRS 等欄位。

結果：

實驗一 test 之 ACC 最高是 LibSVM 及 XGBoost 皆為 64.1%，AUC 最高則是 XGBoost 的 0.641。實驗二 test 之 ACC 最高是 Logistic Regression 為 65.2%，AUC 最高則是 Logistic Regression 的 0.648。實驗三 test 之 ACC 最高是 Random Forest 為 62.6%，AUC 最高則是 XGBoost 的 0.623。

二、引言、背景 Introduction

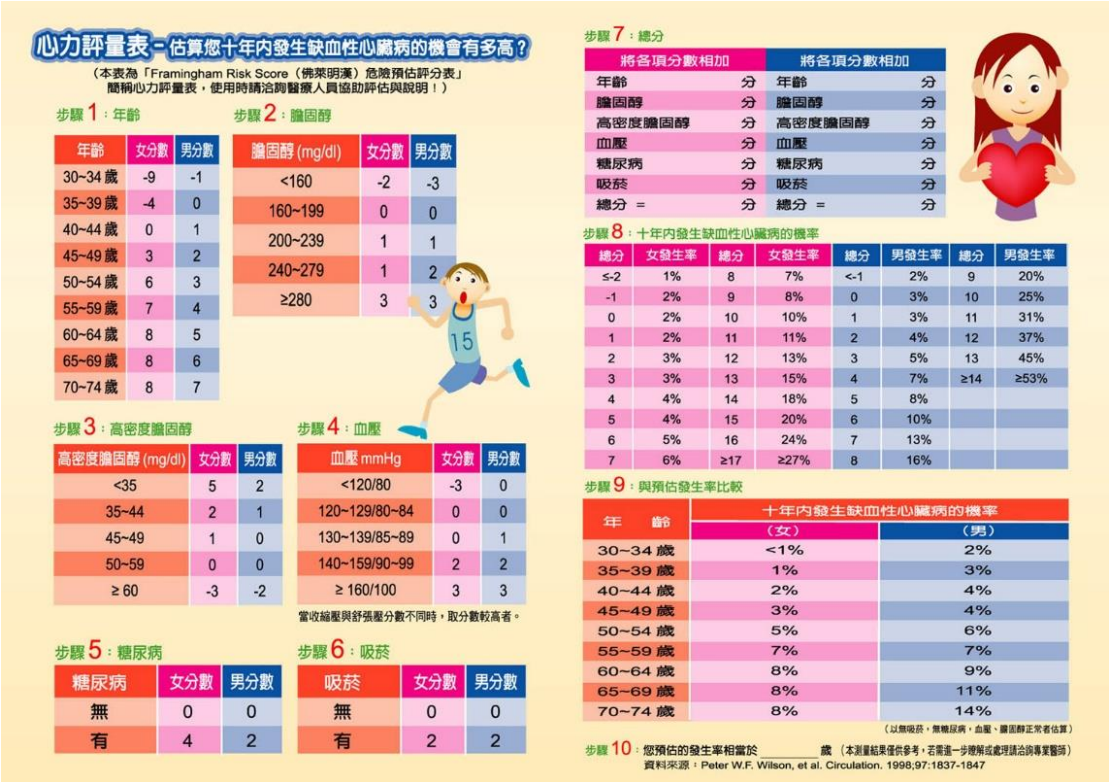
心血管疾病指的是關於心臟或血管的疾病，又稱為循環系統疾病、迴圈系統疾病，其中常見的心血管疾病包括冠狀動脈症候群、中風、高血壓心臟病、周邊動脈阻塞性疾病等等。心血管疾病是全球最常見的死因之一，2013 年造成 1730 萬人死亡，佔死亡人數 31%。根據國內衛生福利部 107 年的死因統計，心臟病是國人第二號殺手，造成 21,569 人死亡，如再加上腦中風、高血壓、糖尿病等血管性疾病，則每年造成 53,977 人死亡，遠超過癌症奪走的人命。由於心血管疾病的高死亡數量，因此所有心血管疾病的成因、防治及治療都是醫學研究的重要範疇，不容輕忽。

冠狀動脈心臟病及成因

冠狀動脈心臟病（簡稱冠心病）是最常見的一種心臟病。所謂冠狀動脈就是位於心臟表面的動脈血管，負責供應心臟血液與養分，一旦血管發生狹窄或堵塞，血液與養分就不易通過，也就會造成心臟缺氧而引發心絞痛，甚至進而導致心臟肌肉壞死，發生急性心肌梗塞、心臟衰竭甚至猝死。

人體血液循環系統中的動脈壁會隨著年齡的增長而形成脂肪堆積（粥狀斑塊），造成動脈狹窄，導致血管阻塞，使血液循環受阻無法暢通流動，醫學上稱之為「動脈粥狀硬化」。其中血管鈣化是動脈硬化的重要指標之一，當血管壁上出現過量鈣的現象時會造成血管動脈硬化及血管狹窄，容易造成心肌梗塞和腦中風等危害。常見造成動脈硬化相關的風險因子包括血中膽固醇異常、高血壓、糖尿病、吸菸、肥胖、動脈粥狀硬化的家族史、以及不健康的飲食習慣。

「十年心血管疾病風險評估」是目前臨床最常使用的評估方法，利用的是弗雷明漢（Framingham）追蹤族群所建立的十年心血管疾病風險預測模型，針對尚未發生因粥狀動脈硬化造成心血管疾病之成年人，利用年齡、抽菸習慣、是否有糖尿病、血壓、血脂肪（總膽固醇、低密度脂蛋白膽固醇、高密度脂蛋白膽固醇）等六項指標，建立心血管風險預測模型，預測出未來十年可能罹患缺血性心臟病的機率，以及心臟年齡的參考值。依照心力評量表可以進行風險推算，將風險分為低度風險、中度風險及高度風險，分別為<10%、10~20%、及>20%。**圖一**為中華民國心臟學會公布的「Framingham Risk Score（佛萊明漢）危險預估評分表」簡稱「心力評量表」：



圖一、Framingham Risk Score（佛萊明漢）危險預估評分表。簡稱心力評量表

冠狀動脈預測的臨床現況及相關研究成果

心血管疾病預測目前的臨床現況較多是利用心臟斷層掃描來得到冠狀動脈鈣化評分，雖然能較精準地預測心血管的風險，但仍然有輻射的相關顧慮以及價格高的缺點。近年來也有許多研究檢測數值，透過機器學習方法預測是否有冠狀動脈心臟病，**表一**為目前所查找到的期刊內容整理。

近年來由於心血管疾病的盛行率及發生率持續增加，該疾病往往沒有明顯的發病前兆，卻留下嚴重的後遺症和遺憾，因此希望能夠利用初階健康檢查項目幫助患者預測冠狀動脈是否狹窄，提供做健康檢查的人當作是否需要高階醫學影像檢查的參考。

	資料數	特徵數	影像	方法	目標	Validation/test	結果
(1)	303(人)	14	X	LR、DT、RF、SVM、KNN、ANN	檢測是否有CAD	test	ANN最好 ACC=0.9303 Recall=0.9380
(2)	6294(人)	15	X	LASSO+ LR、RF、DF、SVM	預測CAD風險	validation	RF的AUC最高 AUC=0.948
(3)	5819(人)	9	X	只寫機器學習方法 (付錢才給看)	建立輔助風險分層系統	10-CV	AUC=0.75
(4)	506(人)	18	X	SVM、KNN、RF、NB、gradient boosting、LR	CAD機器學習預測模型	test	RF➔ACC=92.04% NB➔Spe=92.4% SVM➔Sen=87.34% RF➔ROC=92.20%

表一、預測冠狀動脈心臟病相關研究成果

三. 資料集 Dataset

資料集使用來自彰化基督教醫院 100 年至 108 年對病患的健康檢查原始數據，經前處理及補缺值後，血管狹窄資料共有 4937 筆。

這些資料中有 3111 筆無狹窄(label 0)及 1826 筆有狹窄(label 1)，分別佔資料的 63%和 37%。**圖二**及**圖三**分別為特徵選取的 51 項健檢測量項目及血管狹窄資料基於 Rank-sum test/Pearson Chi test 計算出的特徵顯著性。

性別 Gender	年齡 Age	吸菸 Smoke	喝酒 Alcohol	自述個人病史 PerH	家族病史 FamH	身高 Height
體重 Weight	脈博 Pulse	腰圍 WaistLine	收縮壓 SBP	舒張壓 DBP	脂肪率 PBF	身體質量指數 BMI
心血管系統 Cvsys	白血球 WBC	紅血球 RBC	血色素 HB	血球容積比 HCT	紅血球體積 MCV	平均血紅素 MCH
平均血球血紅素濃度 MCHC	血小板數 PLT	紅血球分佈寬度 RDW	淋巴球 Lymphocyte	單核球 Monocyte	嗜伊紅性白血球 Eosinophil	嗜鹼性白血球 Basophil
紅血球沈降率 ESR	尿蛋白 protein	同半胱胺酸 Homocystein	飯前血糖 Glucose	糖化血色素 HbA1C	血清蛋白 TP	白蛋白 Albumin
白蛋白和球蛋白比 A/G	直接膽紅素Bili D	全膽紅素 Bili T	GOT	GPT	γGT (GGT)	鹼性磷酸胸 ALKP
總膽固醇 TC	高密度膽固醇 HDL	低密度膽固醇 LDL	總膽固醇和HDL比值 TC/HDL	三酸甘油脂 TG	尿酸 Uric acid	尿素氮 BUN
肌酸酐 Creatinine	C-反應蛋白 CRP					

■ : binomial, 二元特徵 [0/1]
■ : categorical, 類別特徵，類別(level)之間沒有大小關係
■ : continuous, 連續特徵，數值之間有大小關係

圖二、特徵選取的 51 項健康檢查測量項目

9.94E-14	7.28E-34	0.17286	8.12E-05	0	0.30183	0.003507
6.36E-08	0.13753	8.62E-13	2.81E-15	0.000192	0.16063	0.000225
0.46572	0.2288	0.19261	0.000451	0.000967	0.001603	0.000576
0.13319	0.001528	0.41592	0.039103	0.000735	0.010592	0.18517
0.22623	0.9894	5.52E-09	7.90E-18	1.82E-20	0.76367	0.46633
0.8714	0.28124	0.23498	0.001213	0.021157	0.006576	0.99013
0.006625	2.85E-11	0.004592	0.000671	9.46E-06	7.75E-07	0.002434
2.54E-07	0.020679					

圖三、血管狹窄資料基於 Rank-sum test/Pearson Chi test 計算出的特徵顯著性
有顏色標記出來的是 p 值<0.05 的特徵

四、方法 Method

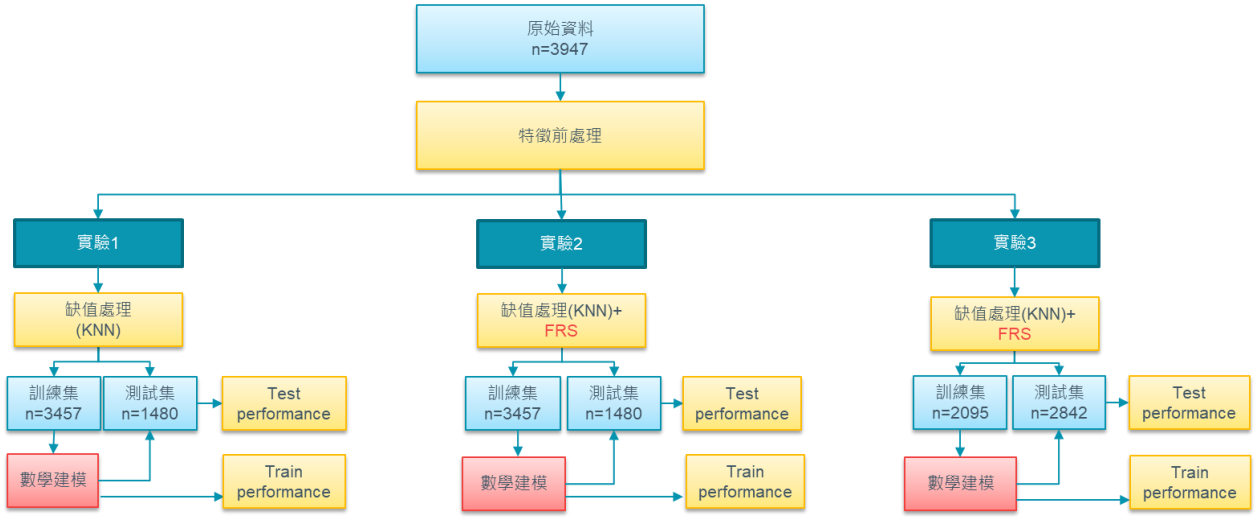
實驗流程（圖四）

將原始資料經特徵前處理後進行三個實驗：

實驗一：將原始的數據經過 KNN 補值，沒有增加佛萊明漢分數（FRS）欄位的數據。

實驗二：上述的檔案增加佛萊明漢分數（FRS）欄位。

實驗三：將原始資料集中沒有缺值的當作訓練集，有缺值的資料經過 KNN 補值後放入測試集，訓練集和測試集同樣都有增加佛萊明漢分數（FRS）欄位。

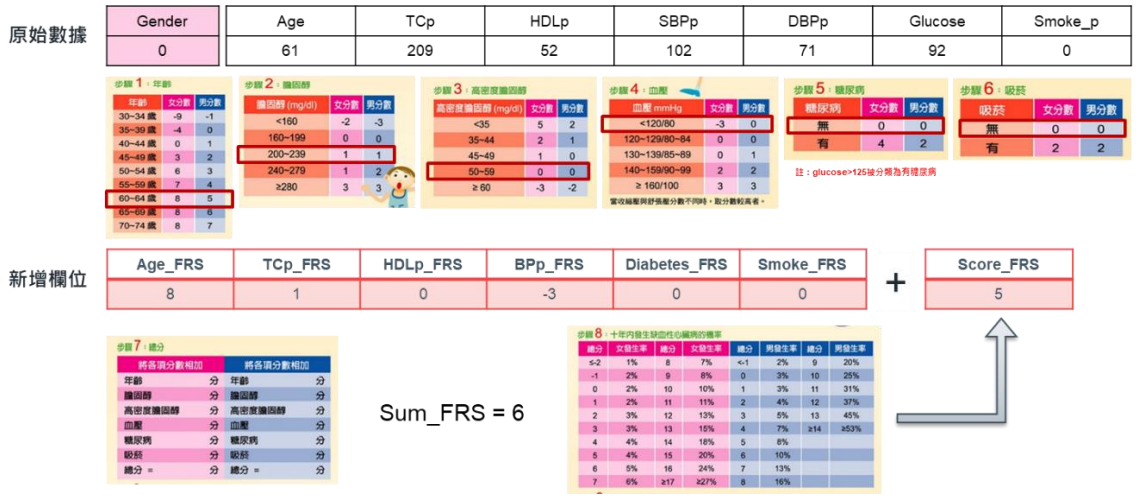


圖四、實驗流程

資料統計(total=4937)				
• 實驗1(Feature:51)：原始data經KNN補值 • 實驗2(Feature:51+6+1)：上述檔案加上FRS	Train (3457)	Label: 0	n=2178	70%
		Label: 1	n=1279	30%
	Test (1480)	Label: 0	n=933	70%
		Label: 1	n=547	30%
• 實驗3(Feature:51+6+1) Train：無缺值➔加上FRS Test：有缺值經KNN補值➔加上FRS	Train (2095)	Label: 0	n=1361	65%
		Label: 1	n=734	35%
	Test (2842)	Label: 0	n=1750	62%
		Label: 1	n=1092	38%

表二、各實驗的資料統計

加入佛萊明漢分數（FRS）欄位



圖五、新增佛萊明漢分數欄位流程表

圖五為利用一位病人的原始數據來增加其佛萊明漢分數（FRS）欄位的流程表。將病人的各類原始數據比對心力評量表的指標，獲得各項目的佛萊明漢分數（Age_FRS、TCp_FRS、HDLp_FRS、BPp_FRS、Diabetes_FRS、Smoke_FRS），最後加總起來與「十年內發生心臟病機率」比對得到 Score_FRS（風險評分）欄位。總欄位為 7（6+1），為 Age_FRS、TCp_FRS、HDLp_FRS、BPp_FRS、Diabetes_FRS、Smoke_FRS 及 Score_FRS。

*註：糖尿病的數據則是根據許多醫院將是否有糖尿病 125 設為分界

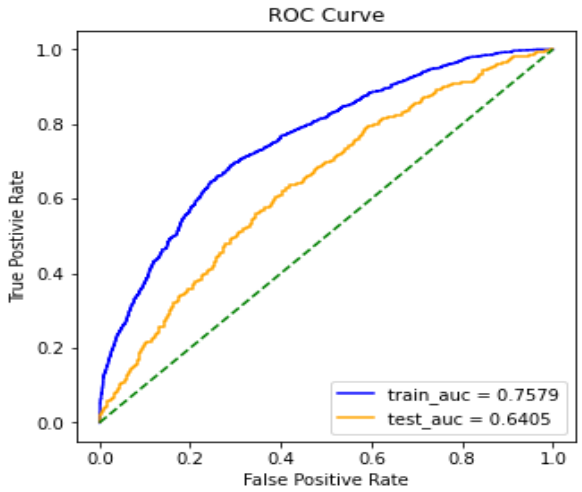
五、結果 Result

三個實驗都使用了 WEKA（SMO、Naïve Bayes、Random Forest、Logistic Regression）、LibSVM、XGBoost 及 IBCGA+SVM 等機器學習方法。

實驗一（經 KNN 補值無 FRS）

		SMO	NB	RF	LR	LibSVM	XGB	IBCGA+SVM
Train	ACC	0.630	0.612	1	0.652	0.693	0.696	0.669
	AUC	0.500	0.615	1	0.660	-	0.758	0.697
10-CV	ACC	0.630	0.606	0.644	0.634	-	-	0.643
	AUC	0.500	0.599	0.625	0.628	-	-	0.635
Test	ACC	0.630	0.622	0.635	0.635	0.641	0.641	0.562
	AUC	0.500	0.626	0.607	0.634	-	0.641	0.608

表三、實驗一數據



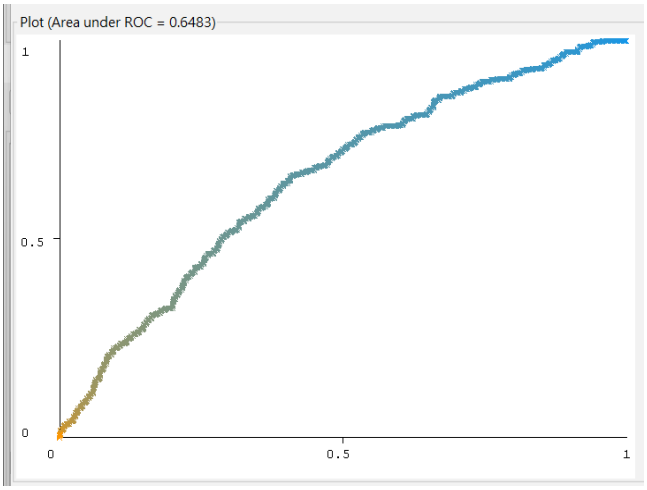
圖六、XGBoost 的 ROC curve，test_auc=0.641

Test 的準確率（ACC）中最高的是 LibSVM 及 XGBoost 為 64.1%，AUC 的話則是 XGBoost 最高為 0.641。圖六為實驗一 XGBoost（AUC 最高）的 ROC Curve。

實驗二（經 KNN 補值有 FRS）

		SMO	NB	RF	LR	LibSVM	XGB	IBCGA+SVM
Train	ACC	0.630	0.619	1	0.642	0.705	0.701	0.641
	AUC	0.500	0.626	1	0.650	-	0.753	0.700
10-CV	ACC	0.630	0.615	0.636	0.640	-	-	0.634
	AUC	0.500	0.614	0.617	0.628	-	-	0.639
Test	ACC	0.630	0.616	0.641	0.652	0.632	0.647	0.565
	AUC	0.500	0.635	0.619	0.648	-	0.645	0.615

表四、實驗二數據



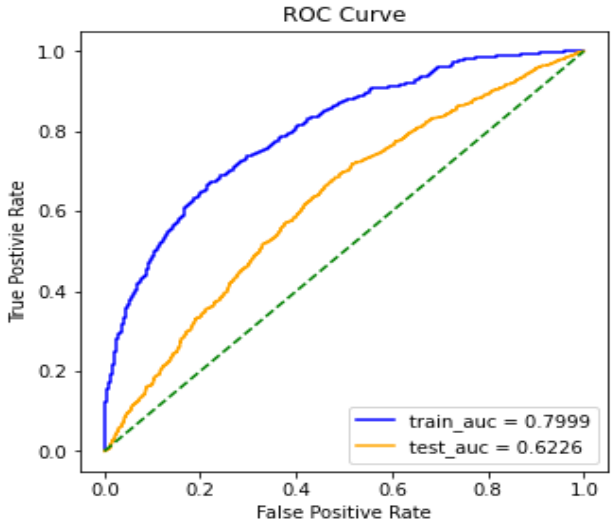
圖七、WEKA-Logistic Regression 的 ROC curve，test_auc=0.648

Test 的準確率（ACC）中最高的是 Logistic Regression 為 65.2%，AUC 的話則是 Logistic Regression 最高為 0.648。**圖七**為實驗二 WEKA-Logistic Regression（AUC 最高）的 ROC Curve。

實驗三（Train：無缺值、Test：有缺值經 KNN 補值有 FRS）

		SMO	NB	RF	LR	LibSVM	XGB	IBCGA+SVM
Train	ACC	0.655	0.639	1	0.681	0.667	0.744	0.700
	AUC	0.528	0.656	1	0.696	-	0.799	0.700
10-CV	ACC	0.643	0.629	0.653	0.657	-	-	0.650
	AUC	0.510	0.641	0.642	0.650	-	-	0.671
Test	ACC	0.612	0.600	0.626	0.620	0.619	0.620	0.567
	AUC	0.513	0.608	0.618	0.607	-	0.623	0.567

表五、實驗三數據



圖八、XGBoost 的 ROC curve，test_auc=0.622

Test的準確率(ACC)中最高的是Random Forest為62.6%，AUC的話則是XGBoost最高為0.622。**圖八**為實驗三 XGBoost（AUC 最高）的 ROC Curve。

IBCGA 特徵挑選

	1(無FRS) (train=3457 / test=1480)	2(有FRS) (train=3457 / test=1480)	3(有FRS-train無缺值) (train=2095 / test=2842)
Feature selection	Age PerH HbA1C SBPp ESR HDLp Creatinine_p HCT UricAcid Gender Weight_p Pulse GPT GGT (14)	Score_FRS Age HbA1C PerH HcT GPT ESR Gender Eosinophil Smoke_FRS SBPp Weight_p Alcohol_p uProtein Creatinine_p GGT plt BPp_fRS (18)	Score_FRS HbA1C Age BMIp PerH HB ESR Height_p Pulse SBPp Smoke_FRS Basophil PBF Smoke_p (14)

表六、IBCGA 特徵挑選出的特徵列表（由主效果大到小）

表六為各實驗經 IBCGA 特徵挑選出的特徵列表，由主效果大到小排序。黃色標記的是 3 個實驗中都有被挑選出來的特徵，而綠色標記的是增加的 FRS 特徵有被選出來的，其中 Score_FRS 和 Smoke_FRS 分別為新增的佛萊明漢風險評分及有無抽菸的分數。

六、討論 Discussion

實驗結果數據

	1(無FRS) (train=3457 / test=1480)				2(有FRS) (train=3457 / test=1480)				3(有FRS-train無缺值) (train=2095 / test=2842)			
	train		test		train		test		train		test	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
SMO	0.630	0.500	0.630	0.500	0.630	0.500	0.630	0.500	0.655	0.528	0.612	0.513
NB	0.612	0.615	0.622	0.626	0.619	0.626	0.616	0.635	0.639	0.656	0.600	0.608
RF	1	1	0.635	0.607	1	1	0.641	0.619	1	1	0.626	0.618
LR	0.652	0.660	0.635	0.634	0.642	0.650	0.652	0.648	0.681	0.696	0.620	0.607
LibSVM	0.693	-	0.641	-	0.705	-	0.632	-	0.667	-	0.619	-
XGboost	0.696	0.758	0.641	0.641	0.701	0.753	0.647	0.645	0.744	0.799	0.620	0.623
IBCGA+SVM	0.669	0.697	0.562	0.608	0.641	0.700	0.565	0.615	0.700	0.700	0.567	0.607

表七、各實驗 train 及 test 之 ACC 和 AUC 數據

表七中的標題列為實驗項目及其訓練集和測試集的樣本數，首欄為分類器。

根據**表七**，以整體趨勢來看，XGboost 和 Logistic Regression 在各個實驗中與其他分類器相比有較好的準確率（ACC）和 AUC。其中實驗 2 的 Logistic Regression 的準確率（ACC）是當中最高且達到 65.2%，其 AUC 也是最高 0.645。單純拿 Logistic Regression 和 XGBoost 比較的話，只有在實驗二 XGBoost 的表現較 Logistic Regression 差，可能的原因是 XGBoost 算是集成式學習的方法，調整參數時很容易會發生 overtraining 的狀況。這兩者在各實驗的表現沒有很明顯的落差，可能可以進行統計方法計算並比較是否真的有顯著差異。

另外，可以看到實驗 2 的 ACC 和 AUC 整題而言都有比實驗 1 的結果好一些，因此推論加入 FRS 分數可能對訓練模型產生幫助。而實驗 3 在大部分的分類器中 ACC 和 AUC 都較低，可能的原因是因為一開始要求訓練集不能有缺值，所以訓練集的樣本數較少（train=2095），導致模型可能訓練得不好。

IBCGA+SVM 的部分可以看到可能有 underfitting 的狀況，未來可以利用 domain knowledge 增加對模型有用或相關的特徵，也能考慮利用集成式學習的方式或結合分類器建成 ensemble model 來增加模型的複雜度。

IBCGA 特徵挑選

	1(無FRS) (train=3457 / test=1480)	2(有FRS) (train=3457 / test=1480)	3(有FRS-train無缺值) (train=2095 / test=2842)
Feature selection	Age PerH HbA1C SBPp ESR HDLp Creatinine_p HCT UricAcid Gender Weight_p Pulse GPT GGT (14)	Score_FRS Age HbA1C PerH HcT GPT ESR Gender Eosinophil Smoke_FRS SBPp Weight_p Alcohol_p uProtein Creatinine_p GGT plt BPp_fRS (18)	Score_FRS HbA1C Age BMIp PerH HB ESR Height_p Pulse SBPp Smoke_FRS Basophil PBF Smoke_p (14)

表六、IBCGA 特徵挑選出的特徵列表（由主效果大到小）

根據**表六**可以看到，IBCGA 特徵挑選中黃色標記的特徵 Age、PerH、HbA1C、SBPp、ESR 是三個實驗中都有被挑選出來的特徵，這些特徵理論上就是在訓練模型時較重要的特徵。綠色標記的是增加的 FRS 特徵有被選出來的是 Score_FRS 和 Smoke_FRS 分別為佛萊明漢風險評分及有無抽菸的分數，也能推測新增的欄位中，這兩個特徵與血管是否有狹窄有較顯著的相關性。

資料來源 References

https://www.cmuh.cmu.edu.tw/HealthEdus/Detail?no=5170
https://www.liver.org.tw/journalView.php?cat=73&sid=1067&page=1
https://www.rsroc.org.tw/knowledge/education/content.asp?ID=47
https://www.cch.org.tw/vmpc/news/news_detail.aspx?oid=224
https://www.frontiersin.org/articles/10.3389/fcvm.2019.00172/full
https://pubmed.ncbi.nlm.nih.gov/30060039/
https://pubmed.ncbi.nlm.nih.gov/32818267/
https://www.future-science.com/doi/10.2144/fsoa-2020-0206
https://www.frontiersin.org/articles/10.3389/fcvm.2021.614204/full
https://www.internationaljournalofcardiology.com/article/S0167-5273(20)33900-0/fulltext
https://link.springer.com/article/10.1007/s42979-021-00731-4

課堂心得：
這堂課應該算是這三年來修習最接近資訊工程相關的課程，雖然跟資訊工程的機器學習來比應該還是很淺，但透過這堂課不僅學習到很多機器學習的方法，也對於這個領域的研究方向更有概念。過程中查找機器學習和深度學習相關期刊和 paper 時也覺得其他領域的研究十分有趣，也發現利用這些機器學習、深度學習等方法來做研究可以在各方面領域幫助這個世界很多，不單單只是運用在智慧醫療方面。

然而除了學到很多之外，從學習過程中也認知到自己在許多方面還有需要加強的地方，不論是在實驗設計、數據分析，甚至是最基本的程式撰寫都深深讓我發現自己還有很多需要加強的面向。一開始想得很遠希望可以做到哪裡，但到後面發現自己在程式撰寫上可能因為距離之前打程式已經有一段時間了，熟悉度以及語法的運用都有些生疏，加上過去沒有學過 panda 跟 numpy，面對 code 以及各種 debug 或是看到了某個 package 結果發現裡面一大堆不懂的參數也有讓我有點心累的時候 XD，多虧這堂課讓我這個學期好幾天都看到太陽升起來，不過最後看到那些數據還有自己劃出的圖表跟 ROC curve 那幾條線的時候還是很開心！

過程雖然對我來說有點辛苦，但是也讓我更加了解所謂做研究也是有花了很多時間卻不一定有好成果的時候，面對數據時常常有很多疑問，也會懷疑自己到底有沒有地方做錯還是跑錯模型什麼的，不過偷偷在這裡 call out 亭君學姊，感謝學姊願意跟我討論還有回答我的問題，一個人看數據的時候常常陷入自我懷疑，而且菜鳥的我面對數據結果以及曲線只能給出偏粗糙的分析，但多虧了學姊願意跟我討論並耐心回答我的問題讓我在面對數據的時候不至於那麼慌張，也能更有自信地面對不同實驗的結果。或許這次書面報告裡面的分析有可能還是沒有那麼詳盡或是精準（是我的問題 QQ），但這次的撰寫書面報告應該也是有讓我加強了數據結果的表達能力，以後可以再多看 paper 加強。最後也要感謝老師跟佩汶助教這個學期的教導和幫助，雖然我還是很菜，但是有跌倒撞壁也有收穫的學習過程整體下來感覺也是挺不錯的！

p.s.我在研究 Lasso 時發現自己還不太了解 panda 跟 numpy，所以最近期末完決定來熟悉 panda 跟 numpy，希望可以幫助到之後的程式撰寫 QQ