科技部補助

大專學生研究計畫研究成果報告

計畫名稱：探討大腸直腸癌倖存者罹患多發性惡性腫瘤及其危險因子的預測

執行計畫學生：蔡婷瑀

學生計畫編號：MOST 106-2813-C-040-003-E
研究期間：106 年 07 月 01 日至 107 年 02 月 28 日止，計 8 個月
指導教授：張啟昌

處理方式：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

執行單位：中山醫學大學醫學資訊學系

中華民國 107 年 03 月 31 日

科技部補助

大專學生研究計畫研究成果報告

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
＊ 計畫 ＊
＊ ：探討大腸直腸癌倖存者罹患多發性惡性腫瘤及其危險因子的預測 ＊
＊ 名稱 ＊
＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

執行計畫學生：蔡婷瑀

學生計畫編號：106-2813-C-040 -003 -E

研 究 期 間 : 2017 年 7 月 1 日至 2018 年 2 月 28 日止,計 8 個月

指 導 教 授 : 張啟昌

處理方式(請勾選)：☐立即公開查詢

☐涉及專利或其他智慧財產權，☐一年 ☑二年後可公開查詢

執 行 單 位：中山醫學大學

中華民國 107 年 3 月 31 日

(一) 摘要

大腸直腸癌(Colorectal Cancer, CRC)在臨床上通常是依據疾病的發展提供適合的進程治療。因此，對於癌症復發徵候的偵測及其後續無症狀復發事件的觀察而言，是與個體的存活率密切相關。過去很多研究將變因的觀察以全民健保資料庫抽樣檔的門診處方及治療明細檔作為資料分析，缺乏實際觀察個別病患深入特定臨床路徑的移轉、復發和治療的時序關聯樣式，以提供臨床醫師對可能的病情發展有更多資訊可參考。因此，為了提高治癒率與存活率，從實際診療紀錄中找出預測復發及產生第二癌因子提供臨床醫師治療的資訊是非常關鍵且重要。本研究所需的病歷記錄和病理資料的來源為五處醫院癌症防治中心癌症登記資料庫。初步經由資深臨床醫師討論並決定第二癌症的危險因子共 20 個變數。經過資料清理後共計有效個案 1223 筆。首先，藉由使用四種分類器：最鄰近搜尋法(IBK)、KSTAR 以及 RandonizableFilteredClassifier 和隨機樹(RandomForest)深入分析敏感度、特異度及預測準確率。整體結果 RandomTree 模式比其他方法有更好的分類準確度。其中，以第二癌為目標變數，RandomTree 準確率為 99.231%最高，最高的危險因子是腫瘤大小(Tumor Size)；以復發為目標變數，RandomTree 準確率最高 99.158%，首位的危險因子是手術邊緣(Surgical Margins)；以複合目標變數：有第二癌及有復發，RandomTree 準確率為 99.534%最高，最重要的危險因子是手術邊緣；以複合目標變數：有第二癌但沒有復發，RandomTree 準確率最高為 99.395%，首位危險因子是腫瘤大小；以複合目標變數：沒有第二癌但有復發，最高的準確率是 IBK (99.052%)，第一位危險因子為手術邊緣；以複合目標變數：沒有第二癌及沒有復發，RandomTree 顯示有最高的準確率 99.256%，其首位危險因子是手術邊緣。整體而言，發生第二癌與復發的兩項共同重要的危險因子是：手術邊緣(Surgical Margins) 和腫瘤大小(Tumor Size)，可經由本研究架構的分層樣式提供臨床醫師輔助治療之參考。

(二)研究動機與研究問題

大腸直腸癌是西方已開發國家前三大癌症死因(Ong et al., 1997)，在台灣也是國人第一常見的癌症死因(衛生福利部國民健康署, 2016)。治療大腸直腸癌的主要方法是外科手術，然而有超過三分之二原發性疾病的病人接受可行的治癒方式，並將所有的腫瘤切除，仍然高達 50%的病患五年內終將死亡(Walker et al., 2014)，其中多數來自於局部、區域性或遠端的腫瘤復發，但由於腸道裡不同位置的腫瘤細胞可能有不同的復發行為模式，所以很難進行預測分析。大腸直腸癌病患在第一期約有 5-10%；第二期約有 20%以及第三期約有 35%會因產生復發導致死亡(Leung and Liu, 2014)，故早期發現和治療非常重要。面對大腸直腸癌症的復發型態，迄今在文獻中還是無法有一致的結論。因此本研究經過文獻與臨床資深醫師討論後，整理復發因素及產生第二癌因素包括 20 項：(1) 年齡(Age) (2) 原發部位(Primary Site) (3)組織型態(Histology) (4)性態碼(Behavior Code) (5)分化(Differentiation) (6)腫瘤大小(Tumor Size) (7)病理期別(Pathologic Stage Group) (8)手術邊緣(Surgical Margins) (9) 手術(Surgical) (10) 放射(RT) (11)手術前放射 (RT surgery) (12)區域治療與全身性治療順序 (Sequence of Locoregional Therapy and Systemic Therapy) (13)最高劑量(Dose to CTV_H) (14)最高次數(Num to CTV_H) (15) 較低劑量(Dose to CTV_L) (16) 較低次數(Num to CTV_L) (17) BMI (18) 吸菸(Smoking) (19) 檳榔(Betel Nut) (20) 喝酒(Drinking)，進一步研究上述潛在危險因子中，分析何者是影響復發及產生第二癌的重要變數。

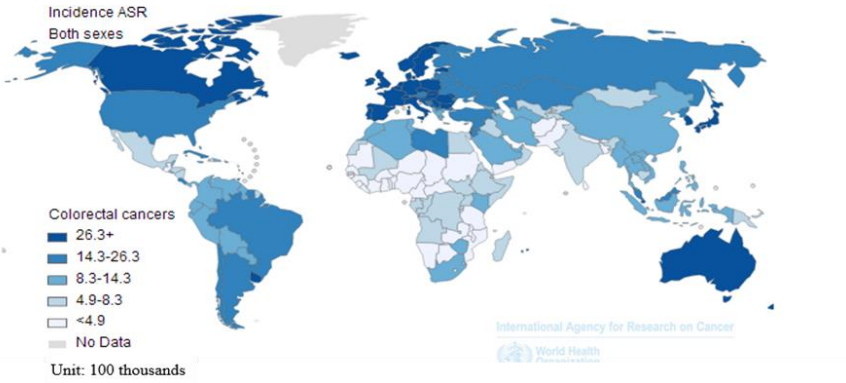
隨著資訊技術的發展，資料探勘（Data Mining）技術逐漸成為臨床診療指引及教學研究上最有價值的工具。所謂的資料探勘又稱之為機器學習（Machine Learning）就是從儲存於資料庫中的資料表、資料記錄及資料欄位內容裡的大量資料中分析出感興趣而隱藏於資料集內的重要資訊。利用資料探勘方法的分類技術也已經成為國內外熱門的研究領域，在此種情況下，使用現代的資料探勘方法可找出大腸直腸癌復發重要因子之間的關聯。

(三)文獻回顧與探討

大腸直腸癌是西方已開發國家前三大癌症死因，在台灣也是國人第一常見的癌症死因。根據 International Agency for Research on Cancer (IARC)最新的資料統計顯示，大腸直腸癌是全球男性第三常見的癌症(746,000 個案例，佔罹患癌症人口的 10%)和全球女性第二常見的癌症(614,000 個案例，佔罹患癌症人口的 9.2%)，將近 55%的案例發生在開發國家。發生率橫跨全世界且在男性與女性間十

分相近：在全世界兩性的發生率相差了 10 倍，最高估計值是澳洲、紐西蘭(44.8%和 32.2%)，而最低為西非(每 10 萬男性與女性年齡標準化率分別為 4.5%和 3.8%)，如圖一所示。

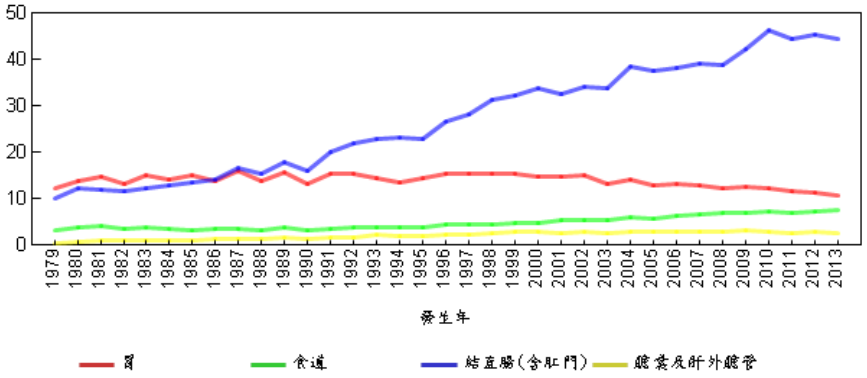
圖一、大腸直腸癌-全球不分性別發生區域



(資料來源：International Agency for Research on Cancer (IARC), 2016)

根據衛生福利部國民健康署的統計，2013 年台灣大腸直腸癌發生率約為每 10 萬人口 15,140 例，平均年齡約在 66.02 歲，而死亡率約為每 10 萬人口 5,265 例，平均年齡約在 70.59 歲，在十大癌症死亡率中占第三位，僅次於肺癌及肝癌。如圖二與表一分別表示胃癌、食道癌、結直腸癌及膽囊及肝外膽管癌在台灣的發生率(1979-2015)。

圖二、胃癌、食道癌、結直腸癌及膽囊及肝外膽管癌在台灣的發生率(1979-2013)



(資料來源： 衛生福利部國民健康署，2016)

隨著癌症治療技術的進步，大幅提升預後與存活率；在此同時，如何謹慎觀察罹患多發性惡性腫瘤(MPMNs)，一直是醫師與病人重視的問題。多發性惡性腫瘤是在個體診斷出在不同組織、器官的兩個或更多的獨立原發惡性腫瘤(Kai et al., 2012; Li et al, 2015)。一般而言，多發性惡性腫瘤以雙重癌較為多見。

台灣近年來由於癌症篩檢的推廣，使得早期發現與診斷成為可行；再加上治療儀器與技術的進步(例如，三度空間順形放射治療、強度調控放射治療、近接放射治療等)。讓更多患者有長的生存時間，但卻形成後續復發與多發性惡性腫瘤的危害(Sakellakis et al., 2014; Santangelo et al., 2015; Xu et al., 2016)。台灣多發性惡性腫瘤的發生頻率正逐年攀升。根據 Hewitt、Lois 等學者(2006)與美國醫學學會(Institute of Medicine)對多發性惡性腫瘤的十大防治建議中第三項指導方針：以癌症登記人口為基礎，運用實證醫學觀點與系統化分析癌症治療技術評估，以建構多發性惡性腫瘤(MPMNs)的臨床治療指引，是刻不容緩需要解決的問題。癌症佔全民健保總支出：由 2006 年(369.5 億)元至 2010 年(553.6 億)元，成長約 1.5 倍。2014 年健保署公布十大癌症醫療支出統計，以大腸直腸癌支出 109 億元台幣最高，其次為肺癌(108 億)及乳癌(103 億)，其餘為肝癌(84 億)、口腔癌(66 億)、白血病(40 億)、非何杰金淋巴癌(39 億)、攝護腺癌(31 億)、胃癌(25 億)、食道癌(23 億)；預估至 2018 年癌症醫療支出會增至 816.4 億元。

本計畫將藉由透過對大腸直腸癌登資料檔的分析，萃取出大腸直腸癌第二癌症防治的有用資訊，進而提供臨床治療的輔助。大腸直腸癌病患是容易得到第二個原發癌症(主要是罹患第二個大腸直腸癌)的高危險族群(9%~28%) (Papaconstantinou et al., 2015)。與一般非癌症患者，大腸直腸癌病患得到第二個大腸直腸癌的風險是 1.5~2.0 倍。其他第二原發惡性腫瘤包括：乳癌、前列腺癌、泌尿器官癌

與肺癌，發生原因主要與遺傳、生活習慣、環境與治療方式有關(Corkum et al., 2013; Sun et al., 2014)。

表一、台灣不分性別每 10 萬人口標準化發生率(2000 年世界標準人口)，1979-2015 年

年度	胃					食道					結直腸癌					膽囊及肝外膽管				
	個案數	平均年齡	年齡中位數	標準化率	百分比%	個案數	平均年齡	年齡中位數	標準化率	百分比%	個案數	平均年齡	年齡中位數	標準化率	百分比%	個案數	平均年齡	年齡中位數	標準化率	百分比%
1979	1,471	57.31	59	12.09	11.33	377	59.23	59	3.16	2.90	1,255	54.11	57	10.09	9.67	64	56.95	58	0.51	0.49
1980	1,734	58.02	60	13.96	11.15	478	59.60	60	3.87	3.07	1,552	55.14	57	12.20	9.98	94	56.32	57	0.75	0.60
1981	1,859	59.04	60	14.84	11.63	503	61.41	61	4.11	3.15	1,550	55.64	57	11.94	9.70	104	60.00	62	0.86	0.65
1982	1,706	59.79	61	13.17	11.15	434	61.94	62	3.38	2.84	1,528	56.41	58	11.48	9.99	107	59.88	63	0.84	0.70
1983	1,985	59.54	61	14.92	11.15	502	62.42	62	3.88	2.82	1,652	57.59	59	12.31	9.28	115	57.60	58	0.87	0.65
1984	1,940	60.61	62	14.18	10.66	480	63.23	63	3.55	2.64	1,819	57.92	60	12.97	9.99	148	61.15	62	1.13	0.81
1985	2,095	61.14	62	14.92	10.80	451	62.52	63	3.14	2.32	1,929	58.57	60	13.35	9.94	155	59.61	61	1.07	0.80
1986	2,026	60.96	63	13.76	10.37	502	62.33	63	3.43	2.57	2,099	58.94	60	14.12	10.74	165	62.55	63	1.14	0.84
1987	2,458	61.09	63	15.96	10.50	516	63.25	64	3.41	2.20	2,578	59.34	61	16.58	11.01	193	60.22	61	1.23	0.82
1988	2,156	61.18	63	13.69	9.50	502	62.03	63	3.21	2.21	2,461	59.69	62	15.51	10.85	204	60.76	63	1.27	0.90
1989	2,495	62.05	65	15.54	9.25	580	63.80	64	3.64	2.15	2,935	59.91	63	17.88	10.88	261	62.23	63	1.61	0.97
1990	2,194	62.36	65	13.09	8.96	522	62.12	63	3.12	2.13	2,687	60.64	63	16.00	10.97	219	62.23	64	1.33	0.89
1991	2,654	62.95	65	15.42	8.79	622	62.72	64	3.62	2.06	3,473	61.05	63	20.06	11.50	297	64.31	66	1.73	0.98
1992	2,785	62.65	65	15.52	8.39	681	63.35	65	3.84	2.05	3,923	61.53	63	21.85	11.81	294	63.87	66	1.67	0.89
1993	2,680	63.65	66	14.56	7.77	673	63.15	64	3.69	1.95	4,220	61.74	64	22.84	12.23	422	65.48	67	2.34	1.22
1994	2,581	63.73	66	13.59	7.23	701	63.90	65	3.74	1.96	4,411	62.50	65	23.17	12.36	372	64.27	66	1.99	1.04
1995	2,849	64.41	67	14.52	7.59	723	62.98	64	3.75	1.93	4,483	62.56	65	22.87	11.95	397	66.05	68	2.06	1.06
1996	3,077	64.78	67	15.26	7.14	881	63.56	64	4.49	2.04	5,346	63.40	66	26.69	12.40	415	65.77	67	2.09	0.96
1997	3,194	65.13	68	15.41	6.77	895	62.87	64	4.35	1.90	5,845	63.66	66	28.26	12.39	478	66.06	68	2.34	1.01
1998	3,291	65.55	68	15.43	6.34	962	62.63	64	4.52	1.85	6,679	63.74	66	31.34	12.86	533	65.79	67	2.53	1.03
1999	3,386	65.74	69	15.33	6.01	1,009	61.69	63	4.58	1.79	7,124	64.16	66	32.42	12.64	594	67.68	69	2.73	1.05
2000	3,351	66.18	69	14.64	5.69	1,082	61.63	62	4.82	1.84	7,668	64.53	67	33.88	13.02	618	68.25	70	2.76	1.05
2001	3,502	66.54	70	14.79	5.82	1,257	61.41	62	5.38	2.09	7,640	64.88	67	32.56	12.69	630	67.38	69	2.69	1.05
2002	3,694	66.76	70	15.12	5.85	1,310	60.83	61	5.47	2.08	8,251	65.12	68	34.07	13.07	720	67.06	68	3.00	1.14
2003	3,360	66.91	70	13.30	5.30	1,356	60.88	61	5.42	2.14	8,391	65.19	68	33.68	13.24	650	67.99	70	2.61	1.03
2004	3,686	66.61	70	14.08	5.19	1,537	60.14	59	6.00	2.16	9,873	65.18	67	38.45	13.90	775	67.46	70	3.00	1.09
2005	3,506	67.43	70	12.93	4.87	1,530	59.42	58	5.77	2.13	9,938	65.56	67	37.56	13.82	748	68.33	69	2.83	1.04
2006	3,683	67.75	70	13.16	4.85	1,766	59.34	57	6.44	2.33	10,524	65.23	67	38.37	13.86	767	68.43	70	2.75	1.01
2007	3,702	67.53	70	12.78	4.64	1,863	59.15	57	6.59	2.34	11,085	65.77	67	39.13	13.90	835	69.52	72	2.88	1.05
2008	3,657	67.75	70	12.21	4.44	2,035	58.98	57	6.98	2.47	11,397	65.98	68	38.90	13.84	872	69.59	71	2.94	1.06
2009	3,890	68.03	70	12.56	4.34	2,076	58.89	57	6.94	2.32	12,769	66.05	67	42.28	14.24	959	69.07	71	3.14	1.07
2010	3,922	68.21	70	12.25	4.22	2,285	58.38	56	7.39	2.46	14,350	65.59	66	46.30	15.43	948	69.48	71	2.98	1.02
2011	3,869	68.10	70	11.73	4.10	2,221	58.70	57	6.99	2.35	14,331	65.76	66	44.53	15.18	856	69.57	71	2.59	0.91
2012	3,824	68.40	70	11.20	3.91	2,382	58.88	57	7.30	2.44	15,072	65.74	66	45.41	15.43	946	68.99	70	2.81	0.97
2013	3,768	68.58	70	10.74	3.80	2,496	58.74	57	7.46	2.52	15,140	66.02	66	44.32	15.27	923	69.89	71	2.63	0.93
2014	3,822	68.80	70	10.51	3.66	2,629	58.91	57	7.68	2.52	15,916	66.01	66	45.14	15.24	1,048	70.33	72	2.87	1.00
2015	3,849	68.70	69	10.28	3.66	2,587	59.22	58	7.36	2.46	15,579	66.07	66	43.01	14.82	1,066	69.62	70	2.85	1.01

(資料來源： 衛生福利部國民健康署，2018)

(四)研究流程與研究方法

台灣癌症登記資料從 1979 年開始蒐集，1996 年由衛生署癌症登記小組進行癌症流行病學資料的收集，目的以提供衛生主管施政及學術機構研究參考(台灣癌症登記，2016)。癌症死亡率始終位居十大死因之首，投注於癌症研究資源，實屬必要。台灣癌症登記資料庫的研究雖有其限制，但與耗時臨床試驗相比，成本較為低廉。台灣現有的癌症研究資料庫發展條件皆已具備，且蒐集資料並不限於老年人口，具有普遍性的特質，相較於美國癌症登記(SEER)在學術研究發展上具有相當的優勢。因此，本計畫整合中山醫學大學附設醫院、亞東紀念醫院與大里仁愛醫院之癌症防治登記中心資料庫為研究對象。相關資料數據如表二所示。對於第一原發癌與第二原發癌關聯分析和預測問題，現今國內的研究比較少。

表二、研究合作醫院之數據估計量

合作醫院名稱	98 年	99 年	100 年	101 年	102 年
徐元智先生醫藥基金會亞東紀念醫院	2,040	2,032	1,970	1,960	1,951
中山醫學大學附設醫院	1,581	1,636	1,540	1,606	1,623
仁愛醫療財團法人大里仁愛醫院	608	645	653	632	568

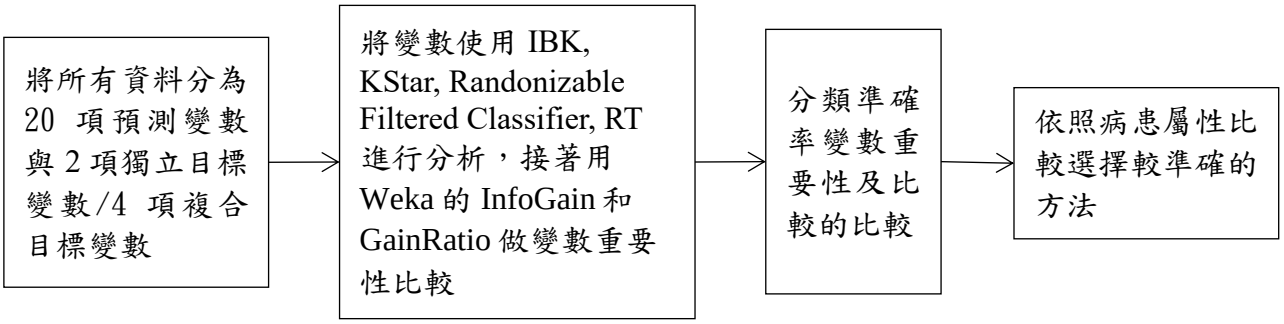
(資料來源：國民健康署癌症登記工作雙季報，2015.6)

隨著資訊技術的發展，資料探勘(Data Mining)技術已經成為臨床診療指引及研究最有價值的工具。資料探勘中運用人工智慧方法最有效率的方法是機器學習(Machine Learning)。機器學習就是從儲存於資料庫內容裡的大量資料分析出感興趣而隱藏資料集內的重要資訊。利用資料探勘方法的分類技術已經成為國內外熱門的研究領域，使用現代的資料探勘方法可找出發生第二癌症重要因子之間的關聯。

本研究以 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval、最鄰近搜尋法(IBK)、KSTAR、RandonizableFilteredClassifier 和隨機樹(RT)五種模型等相關研究基礎，建立大腸直腸癌復發及產生第二癌的重要因子，並探討四種資料探勘方法預測之準確度。

研究流程:

為了比較重要變數篩選的差異，研究設計架構如圖三所示：在圖三中，首先依據文獻查證與臨床醫師討論後決定 20 項預測變數((1)年齡(Age) (2)原發部位(Primary Site) (3)組織型態(Histology) (4)性態碼(Behavior Code)(5)分化(Differentiation) (6)腫瘤大小(Tumor Size) (7)病理期別(Pathologic Stage Group) (8)手術邊緣(Surgical Margins) (9) 手術(Surgical) (10) 放射(RT) (11)手術前放射 (RT surgery) (12)區域治療與全身性治療順序 (Sequence of Locoregional Therapy and Systemic Therapy) (13)最高劑量(Dose to CTV_H) (14)最高次數(Num to CTV_H) (15) 較低劑量(Dose to CTV_L) (16) 較低次數(Num to CTV_L) (17) BMI (18) 吸菸(Smoking) (19) 檳榔(Betel Nut) (20) 喝酒(Drinking)) 進行復發的預測。圖三為第一階段研究流程：直接以 IBK、KSTAR、RandonizableFilteredClassifier 和 RT 方法進行預測，接著以 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 做變數重要性比較。最後，進一步比較方法所分析分類準確率，針對所分析的變數結果，依照病患屬性完成臨床後續預測大腸直腸癌復發及產生第二癌之重要因子的建議以及方法之比較。



圖三、研究流程圖

研究方法：

在醫學衛生領域中，資料探勘應用已大幅度地被用來直接取得預測不同群體之間患者的相關資訊。然而，探勘方法分類技術尚未被利用於分析大腸直腸癌復發。因此，本研究試圖利用五種資料探勘方法由大腸直腸癌的資料庫中進行分類並進一步分析集成學習架構的優勢。

一、Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval

GainRatioAttributeEval：通過測量相對於類的增益比來評估屬性的價值。計算過程如下：

$$GainR（Class，Attribute）=（H（Class）-H（Class\,|\,Attribute））/H（Attribute）。\quad (1)$$

InfoGainAttributeEval：通過測量相對於該類別的信息增益的計算結果的屬性的值。計算過程如下：

$$InfoGain（Class，Attribute）= H（Class）- H（Class\,|\,Attribute）。\quad (2)$$

二、最鄰近搜尋法(IBK)

Lazy Learning 的方法在訓練是僅僅是保存樣本集的信息，直到測試樣本到達時才進行分類決策。也就是說這個決策模型是在測試樣本到來以後才生成的。相對與其它的分類算法來說，這類的分類算法可以根據每個測試樣本的樣本信息來學習模型，這樣的學習模型可能更好的擬合局部的樣本特性。kNN 算法的思路非常簡單直觀：如果一個樣本在特徵空間中的 k 個最相似(即特徵空間中最鄰近) 的樣本中的大多數屬於某一個類別，則該樣本也屬於這個類別。其基本原理是在測試樣本到達的時候尋找到測試樣本的 k 臨近的樣本，然後選擇這些鄰居樣本的類別最集中的一種作為測試樣本的類別。在 weka 中關於 kNN 的算法有兩個，分別是 IB1，IBk，IBk 是通過它周圍的 k 個鄰居來判斷測試樣本的類別

在樣本中有比較多的噪音點是（noisy points）時，通過一個鄰居的效果很顯然會差一些，因為出現誤差的情況會比較多。這種情況下，IBk 就成了一個較優的選項了。這個時候有出現了一個問題，k 這個值如何確定，一般來說這個 k 是通過經驗來判斷的。

三、最近鄰居法(KStar)

K *是基於實例的分類器，即測試實例的類基於類似於它的那些訓練實例的類，如由一些相似性函數所確定的。它不同於其他基於實例的學習者，因為它使用基於熵的距離函數。

四、 RandonizableFilteredClassifier

用於對已通過任意過濾器傳遞的數據運行任意分類器的類。 與分類器一樣，過濾器的結構僅基於訓練數據，測試實例將由過濾器處理而不改變其結構。

五、RT

在電腦科學和數學裡面，一個隨機樹是一個經由隨機過程建立的樹或者樹狀圖(arborescence)，而其中有隨機森林演算法，隨機森林演算法是將多數類樣本劃分為數個獨立的子集合；再將每一個獨立子集合進行交叉組合以構成不同的訓練樣本集，並針對不同的訓練樣本集利用決策樹分類器加以學習；最後根據平均加權法產成隨機森林，進而獲得決策規則(吳華芹，2013)。計算方法為給定 K 個分類器以及隨機向量 x、y，定義邊際函數如下：(張華偉等人，2006)

$$-max_{j\neq y}\mathrm{a}v_k I(h_k(\mathrm{mg(x,y)}=\mathrm{a}v_k I(h_k(x)=y)x)=j) \quad (3)$$

其中，I()是可能性函數，邊際函數顯示向量 X 所得到正確分類 y 的平均得票數超過其它任何類平均得票數的程度。由此可知邊際越大分類的可信度就越高。分類器誤差定義：

$$PE^{*}=P_{x,y}(\mathrm{mg(x,y)}<0)$$

將上面的結論推廣到隨機森林函數： $h_k(X)=h(X,\theta_k)$

邊際函數如下：

$$\mathrm{mr}(\mathrm{x},\mathrm{y}) - P_\theta(h(x,\theta)=y) - max_{j\neq Y}P_\theta(h(x,\theta)=j) \tag{4}$$

隨著樹的數目增加，PE^*就會趨向於

$$\mathrm{P}_{x,y}(P_\theta(h(x,\theta)=y) - max_{j\neq Y}P_\theta(h(x,\theta)=j) < 0) \tag{5}$$

而分類器{h(X,θ)}的強度可以表示為

$$\mathrm{s} = \mathrm{E}_{X,Y}mr(x,y) \tag{6}$$

假設 $s \geq 0$，根據契比雪夫不等式，(16) (17) 兩式可以得到：

$$PE^* \leq (\mathrm{var}(\mathrm{mr})) \,/\, \mathrm{s}^2 \tag{7}$$

根據 Breiman(2001)可推導出

$$\begin{aligned}\mathrm{var}(\mathrm{mr}) &= \bar{\rho}\big(E_\theta sd(\theta)\big)^2 \\ &\leq \bar{\rho}E_\theta var(\theta)\end{aligned}$$

$$\geq 1 - s^2 \tag{8}$$

隨機森林的目標誤差上界是$PE^* \leq \bar{\rho}(1-S^2)/S^2$

首先，使用資料探勘軟體 Weka 中的 GainRaitoAttributeEval 與 InfoGainAttributeEval 來篩選重要變數。其次，對癌症登記資料庫各欄位與惡性腫瘤關係研究的評估方法採用平均準確率(Accuracy)與 F1 評分(F1 Score)作為評估結果。對於分類器或者分類演算法，評估指標主要有敏感度(Sensitivity, TPR)、特異度(Specificity)、平均準確率(Accuracy)與 F1 評分(F1 Score)，敏感度是實際為陽性的樣本中，判斷為陽性的比例；特異度是實際為陰性的樣本中，判斷為陰性的比例；平均準確率代表預測正確的準確度；F1 評分是敏感度(Recall)與準確率(Precision)兩者的綜合，F1 評分越高，說明分類模型越穩健。各算式具體定義如下：

	P	N
Y	True Positives (TP)	False Positives (FP)
N	False Negatives (FN)	True Negatives (TN)

$$\text{Specificity} = \frac{TN}{TP+FN} \tag{9}$$

$$TPR = \frac{TP}{TP+FN} \tag{10}$$

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{13}$$

$$F_1 = \frac{2}{1/Precision + 1/Recall} \tag{14}$$

(五) 實證研究

在研究中，我們由中山醫學大學附設醫院、亞東紀念醫院與大里仁愛醫院之癌症防治登記中心資料庫提供的大腸直腸癌數據集，使用 IBK、KSTAR、RandonizableFilteredClassifier 和 RandomTree 驗證其敏感度與特異度，並用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 預測大腸直腸癌復發和產生第二癌之重要因子。數據集中共包含 20 個預測變數，分別為(1) 年齡(Age) (2)原發部位(Primary Site) (3)組織型態(Histology) (4)性態碼(Behavior Code)(5)分化(Differentiation) (6)腫瘤大小(Tumor Size) (7)病理期別(Pathologic Stage Group) (8)手術邊緣(Surgical Margins) (9) 手術(Surgical) (10) 放射(RT) (11)手術前放射 (RT surgery) (12)區域治療與全身性治療順序 (Sequence of Locoregional Therapy and Systemic Therapy) (13)最高劑量(Dose to CTV_H) (14)最高次數(Num to CTV_H) (15) 較低劑量(Dose to CTV_L) (16) 較低次數(Num to CTV_L) (17) BMI (18) 吸菸(Smoking) (19) 檳榔(Betel Nut) (20) 喝酒(Drinking)以及 6 個目標變數為癌序(SPM)、復發型態（Type of Recurrence）、有產生第二癌及復發、有產生第二癌但沒有復發、沒有復發但產生第二癌還有沒有復發也沒有產生第二癌，共 1223 筆資料，隨機選取 367 筆資料為測試樣本，其餘 856 筆資料為訓練樣本，進行重複取樣十次。

5.1. 第二癌預測分析結果

以癌序(SPM)為目標變數，{1}代表：沒有第 2 癌；{2}則代表：有第 2 癌。因此{1-1}(敏感度)代表：原始的判定有第 2 癌，而經由模式判定後亦為有第 2 癌；而{2-2}(特異度)則表示：原始判定為沒有第 2 癌，經由模式判定亦為沒有第 2 癌。

由表 3 知 IBK 的整體正確判別率為 99.092%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為 99.03%：即原始群體為第 1 類的樣本正確的被判別到第 1 類的比率為 99.03%。其中有 4 個原本群體為第 1 類的樣本，被錯分為第 2 類的群體中；而有 3 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表3、使用IBK分類結果

	分類	
類別	1 (沒有第 2 癌)	2 (有第 2 癌)
1(沒有第 2 癌)	363(99.03%)	4(0.97%)
2 (有第 2 癌)	3(0.76%)	364(99.24%)

平均分類準確率：99.092%

由表4可知KSTAR的整體正確判別率為98.414%，而個別的判別正確率以{2-2}(特異度)的比率最高，為98.60%：即原始群體為第2類的樣本正確的被判別到第2類的比率為98.60%；而{1-1}(敏感度)的判別正確率較差，為98.38%。

表4、使用KSTAR分類結果

	分類	
類別	1 (沒有第 2 癌)	2 (有第 2 癌)
1 (沒有第 2 癌)	361(98.38%)	6(1.62%)
2 (有第 2 癌)	5(1.40%)	362(98.60%)

平均分類準確率：98.414%

由表5可知RandonizableFilteredClassifier的整體正確判別率為93.598%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為93.68%：即原始群體為第1類的樣本正確的被判別到第1類的比率為93.68%；而{2-2}(特異度)的判別正確率為79.61%。

表5、使用RandonizableFilteredClassifier分類結果

	分類	
類別	1(沒有第 2 癌)	2 (有第 2 癌)
1 (沒有第 2 癌)	344(93.68%)	23(6.32%)
2 (有第 2 癌)	75(20.39%)	292(79.61%)

平均分類準確率：93.598%

由表6可知RandomTree的整體正確判別率為99.231%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為99.49%：即原始群體為第1類的樣本正確的被判別到第1類的比率為99.49%；而{2-2}(特異度)的判別為97.53%。

表6、使用RandomTree分類結果

	分類	
類別	1(沒有第 2 癌)	2 (有第 2 癌)
1 (沒有第 2 癌)	365(99.49%)	2(0.51%)
2 (有第 2 癌)	8(2.47%)	358(97.53%)

平均分類準確率：99.231%

從表 7 中，我們可以觀察到以癌序(SPM)為目標變數，RandomTree 模式在{1-1}(敏感度)產生最高平均分類準確率，為 99.49%；而 IBK 在{2-2}(特異度)產生最高平均分類準確率，為 99.24%；在整體情況下，我們可以看到 IBK 模式優於 KSTAR、RandonizableFilteredClassifier 和 RandomTree 模式，這表明 IBK 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

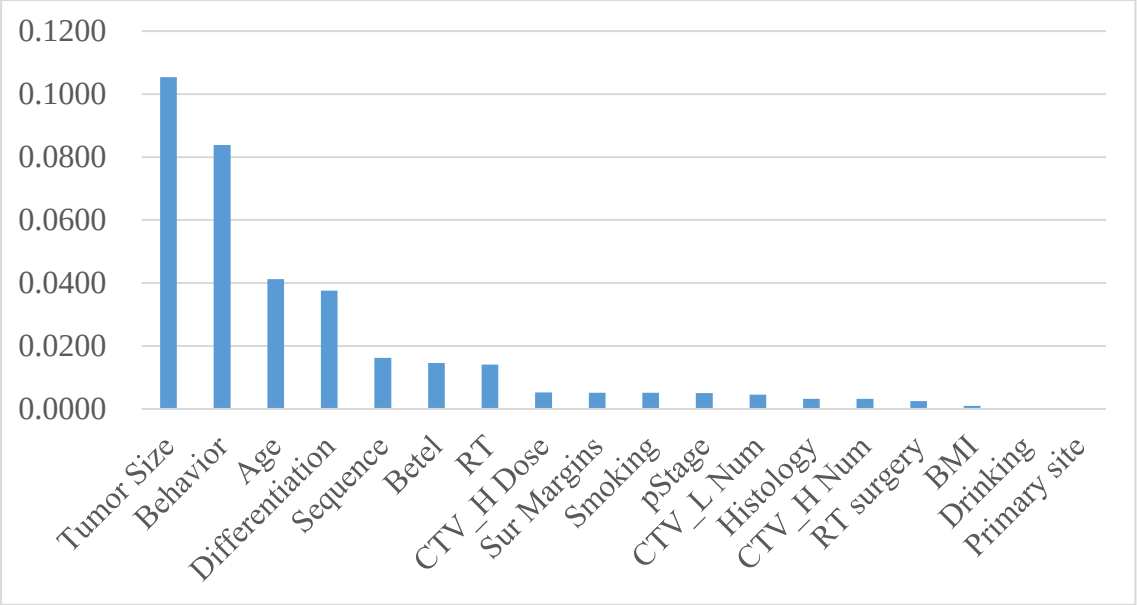
表7、IBK、KSTAR、RandonizableFilteredClassifier(RFC)和RandomTree(RT) 模式預測評估

	預測沒有第 2 癌，實際沒有第 2 癌（敏感度%）{1-1}				預測有第 2 癌，實際有第 2 癌（特異度%）{2-2}				整體平均預測準確率(%)			
模組	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT
1	98.98	98.43	97.94	99.25	99.32	99.29	86.54	96.75	99.02	98.53	96.48	98.94
2	98.67	98.62	75.97	100.0	99.16	98.70	96.26	99.81	99.10	98.69	93.70	99.84
3	98.98	98.15	94.88	99.72	99.32	98.56	78.29	98.09	99.02	98.20	93.13	99.51
4	98.89	98.34	95.20	99.25	99.32	99.29	75.54	98.68	98.94	98.45	92.97	99.18
5	99.25	98.61	95.57	99.44	98.68	98.61	77.86	96.18	99.18	98.61	93.54	99.02
6	99.07	98.15	96.23	98.97	99.32	98.56	85.29	96.05	99.10	98.20	95.01	98.61
7	98.98	98.43	95.73	99.72	99.32	97.90	76.03	97.47	99.02	98.36	93.38	99.43
8	99.25	98.25	95.66	99.72	99.33	98.57	78.01	98.09	99.26	98.28	93.62	99.51
9	98.98	98.34	94.55	99.34	99.32	98.58	69.50	96.77	99.02	98.36	91.66	99.02
10	99.25	98.52	95.09	99.53	99.33	97.92	72.73	97.44	99.26	98.45	92.48	99.26

接著，我們使用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 分析後發現對於產生第二癌的重要變數以腫瘤大小(Tumor Size) 影響最大為 10.53%，原發部位(Primary Site)為最沒有影響力，如表 8 及圖四所示。

表8、Weka的InfoGain和GainRatio針對產生第二癌的重要變數的重要性排序

Gain Ratio			Info Gain			平均重要變數之重要性排序		
重要性排序	百分比	重要變數	重要性排序	百分比	重要變數	重要性排序	百分比	重要變數
1	8.66	Tumor Size (腫瘤大小)	1	2.35	Differentiation (分化)	1	5.27	Tumor Size (腫瘤大小)
2	6.35	Behavior Code (性態碼)	2	2.03	Behavior Code (性態碼)	2	4.19	Behavior Code (性態碼)
3	3.33	Age (年齡)	3	1.87	Tumor Size (腫瘤大小)	3	2.06	Age (年齡)
4	1.40	Differentiation (分化)	4	0.80	Sequence (區域治療與全身性治療順序)	4	1.88	Differentiation (分化)
5	0.99	Betel (檳榔)	5	0.79	Age (年齡)	5	0.82	Sequence (區域治療與全身性治療順序)
6	0.82	Sequence (區域治療與全身性治療順序)	6	0.69	Surgical (手術)	6	0.73	Surgical (手術)
7	0.80	CTV_L Dose (較低劑量)	7	0.67	RT (放射)	7	0.73	Betel (檳榔)
8	0.77	Surgical (手術)	8	0.46	Betel (檳榔)	8	0.70	RT (放射)
9	0.73	RT (放射)	9	0.43	CTV_L Dose (較低劑量)	9	0.62	CTV_L Dose (較低劑量)
10	0.29	Sur Margins (手術邊緣)	10	0.25	pStage (病理期別)	10	0.26	CTV_H Dose (最高劑量)
11	0.29	CTV_L Num (較低次數)	11	0.25	Smoking (吸菸)	11	0.26	Sur Margins (手術邊緣)
12	0.29	CTV_H Dose (最高劑量)	12	0.23	CTV_H Dose (最高劑量)	12	0.26	Smoking (吸菸)
13	0.27	Smoking (吸菸)	13	0.22	Sur Margins (手術邊緣)	13	0.26	pStage (病理期別)
14	0.26	pStage (病理期別)	14	0.15	CTV_L Num (較低次數)	14	0.23	CTV_L Num (較低次數)
15	0.25	Histology (組織型態)	15	0.14	CTV_H Num (最高次數)	15	0.16	Histology (組織型態)
16	0.21	RT surgery (手術前放射)	16	0.07	Histology (組織型態)	16	0.16	CTV_H Num (最高次數)
17	0.18	CTV_H Num (最高次數)	17	0.05	BMI	17	0.13	RT surgery (手術前放射)
18	0.05	BMI	18	0.05	RT surgery (手術前放射)	18	0.05	BMI
19	0.01	Drinking (喝酒)	19	0.01	Drinking (喝酒)	19	0.02	Drinking (喝酒)
20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)



圖四、針對產生第二癌的重要變數的重要性排序直方圖

5.2. 復發預測分析結果

以復發型態（Type of Recurrence）為目標變數，{1}代表：沒有復發；{2}則代表：有復發。因此{1-1}(敏感度)代表：原始的判定沒有復發，而經由模式判定後亦為沒有復發；而{2-2}(特異度)則表示：原始判定為有復發，經由模式判定亦為有復發。

由表 9 知 IBK 的整體正確判別率為 98.921%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為 98.90%：即原始群體為第 1 類的樣本正確的被判別到第 1 類的比率為 98.90%。其中有 4 個原本群體為第 1 類的樣本，被錯分為第 2 類的群體中；而有 5 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表9、使用IBK分類結果

	分類	
類別	1 (沒有復發)	2 (有復發)
1(沒有復發)	363(98.90%)	4(1.10%)
2 (有復發)	5(1.32%)	362(98.68%)

平均分類準確率：98.921%

由表10可知KSTAR的整體正確判別率為98.659%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為98.68%：即原始群體為第1類的樣本正確的被判別到第1類的比率為98.68%；而{2-2}(特異度)的判別正確率較差，為98.45%。

表10、使用KSTAR分類結果

	分類	
類別	1 (沒有復發)	2 (有復發)
1 (沒有復發)	362(98.68%)	5(1.32%)
2 (有復發)	6(1.55%)	361(98.45%)

平均分類準確率：98.659%

由表11可知RandonizableFilteredClassifier的整體正確判別率為91.922%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為91.57%：即原始群體為第1類的樣本正確的被判別到第1類的比率為91.57%；而{2-2}(特異度)的判別正確率為89.97%。

表11、使用RandonizableFilteredClassifier分類結果

	分類	
類別	1(沒有復發)	2 (有復發)
1 (沒有復發)	336(91.57%)	31(8.43%)
2 (有復發)	37(10.03%)	330(89.97%)

平均分類準確率：93.598%

由表12可知RandomTree的整體正確判別率為99.158%，而個別的判別正確率以{2-2}(特異度)的比率最高，為99.07%：即原始群體為第2類的樣本正確的被判別到第2類的比率為99.07%；而{1-1}(敏感度)的判別正確率為98.99%。

表12、使用RandomTree分類結果

	分類	
類別	1(沒有復發)	2 (有復發)
1 (沒有復發)	363(99.07%)	4(0.93%)
2 (有復發)	3(1.01%)	364(98.99%)

平均分類準確率：99.158%

從表 13 中，我們可以觀察到以復發型態（Type of Recurrence）為目標變數，RandomTree 模式在{1-1}(敏感度)產生最高平均分類準確率，為 99.07%；而 RandomTree 也在{2-2}(特異度)產生最高平均分類準確率，為 98.99%；在整體情況下，我們可以看到 RandomTree 模式優於 IBK、KSTAR 和 RandonizableFilteredClassifier 模式，這表明 RandomTree 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

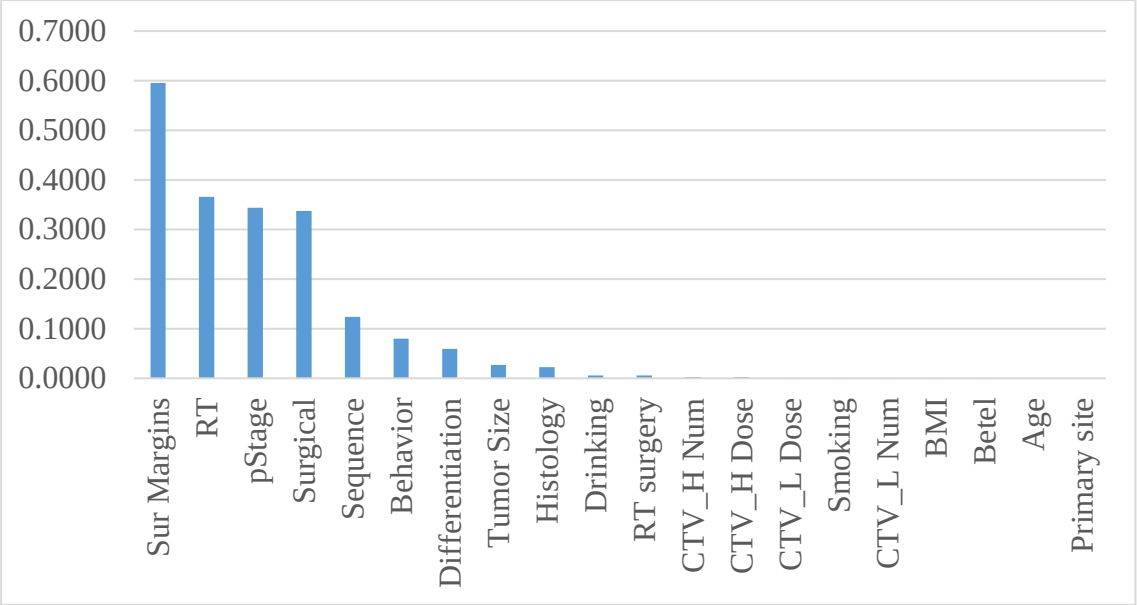
表13、IBK、KSTAR、RandonizableFilteredClassifier(RFC)和RandomTree(RT) 模式預測評估

	預測沒有復發，實際沒有復發（敏感度%）{1-1}				預測有復發，實際有復發（特異度%）{2-2}				整體平均預測準確率(%)			
模組	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT
1	99.18	98.37	87.71	96.24	99.07	98.95	92.78	98.59	99.10	98.77	91.33	97.87
2	98.61	98.37	93.29	99.76	98.90	98.07	87.15	98.40	98.69	98.28	91.50	99.35
3	98.84	98.83	95.24	99.88	98.90	98.36	90.88	99.46	98.86	98.69	93.95	99.75
4	98.84	98.95	95.12	99.30	99.17	98.37	90.36	99.73	98.94	98.77	93.70	99.43
5	98.91	98.35	83.28	98.11	99.18	98.49	90.25	99.18	99.10	98.45	88.31	98.86
6	99.07	98.95	93.34	99.42	99.18	98.63	88.64	99.18	99.10	98.86	91.99	99.35
7	98.95	98.83	94.10	99.53	98.90	98.09	88.86	99.19	98.94	98.61	92.56	99.43
8	98.72	98.95	93.96	99.65	99.17	98.63	87.85	99.46	98.86	98.86	92.15	99.59
9	99.17	98.36	84.96	98.63	98.72	98.83	92.48	98.83	98.86	98.69	90.27	98.77
10	98.72	98.83	94.68	99.41	98.90	98.09	90.50	98.65	98.77	98.61	93.46	99.18

我們使用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 分析後發現對於復發的重要變數以手術邊緣(Surgical Margins)影響最大為 59.57%，原發部位(Primary Site)為最沒有影響力，如表 14 及圖五所示。

表14、Weka的InfoGain和GainRatio針對復發的重要變數的重要性排序

Gain Ratio			Info Gain			平均重要變數之重要性排序		
重要性排序	百分比	重要變數	重要性排序	百分比	重要變數	重要性排序	百分比	重要變數
1	34.23	Sur Margins (手術邊緣)	1	25.34	Sur Margins (手術邊緣)	1	29.79	Sur Margins (手術邊緣)
2	19.09	RT (放射)	2	17.49	RT (放射)	2	18.29	RT (放射)
3	17.72	Surgical (手術)	3	17.10	pStage (病理期別)	3	17.20	pStage (病理期別)
4	17.29	pStage (病理期別)	4	16.00	Surgical (手術)	4	16.86	Surgical (手術)
5	6.27	Sequence (區域治療與全身性治療順序)	5	6.13	Sequence (區域治療與全身性治療順序)	5	6.21	Sequence (區域治療與全身性治療順序)
6	6.04	Behavior Code (性態碼)	6	3.71	Differentiation (分化)	6	3.98	Behavior Code (性態碼)
7	2.24	Tumor Size (腫瘤大小)	7	1.93	Behavior Code (性態碼)	7	2.96	Differentiation (分化)
8	2.21	Differentiation (分化)	8	0.48	Tumor Size (腫瘤大小)	8	1.36	Tumor Size (腫瘤大小)
9	1.79	Histology (組織型態)	9	0.46	Histology (組織型態)	9	1.13	Histology (組織型態)
10	0.45	RT surgery (手術前放射)	10	0.26	Drinking (喝酒)	10	0.28	Drinking (喝酒)
11	0.30	Drinking (喝酒)	11	0.10	RT surgery (手術前放射)	11	0.28	RT surgery (手術前放射)
12	0.13	CTV_H Dose (最高劑量)	12	0.10	CTV_H Num (最高次數)	12	0.12	CTV_H Num (最高次數)
13	0.13	CTV_H Num (最高次數)	13	0.10	CTV_H Dose (最高劑量)	13	0.12	CTV_H Dose (最高劑量)
14	0.03	CTV_L Dose (較低劑量)	14	0.02	Smoking (吸菸)	14	0.02	CTV_L Dose (較低劑量)
15	0.02	Smoking (吸菸)	15	0.02	CTV_L Dose (較低劑量)	15	0.02	Smoking (吸菸)
16	0.01	CTV_L Num (較低次數)	16	0.01	BMI	16	0.01	CTV_L Num (較低次數)
17	0.01	BMI	17	0.01	CTV_L Num (較低次數)	17	0.01	BMI
18	0.00	Betel (檳榔)	18	0.00	Betel (檳榔)	18	0.00	Betel (檳榔)
19	0.00	Age (年齡)	19	0.00	Age (年齡)	19	0.00	Age (年齡)
20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)



圖五、針對復發的重要變數的重要性排序直方圖

5.3. 有第二癌及有復發預測分析結果

以有產生第二癌及有復發為目標變數，{1}代表：非有產生第二癌及有復發的其他 3 種情況；{2}則代表：有產生第二癌及有復發。因此{1-1}(敏感度)代表：原始的判定為其他 3 種情況，而經由模式判定後亦為其他 3 種情況；而{2-2}(特異度)則表示：原始判定為有產生第二癌及有復發，經由模式判定亦為有產生第二癌及有復發。

由表 15 知 IBK 的整體正確判別率為 99.411%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為 99.46%：即原始群體為第 1 類的樣本正確的被判別到第 1 類的比率為 99.46%。其中有 4 個原本群體為第 1 類的樣本，被錯分為第 2 類的群體中；而有 5 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表15、使用IBK分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	365(99.46%)	2(0.54%)
2 (有產生第二癌及有復發)	9(2.48%)	358(97.52%)

平均分類準確率：99.411%

由表16可知KSTAR的整體正確判別率為99.043%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為99.15%：即原始群體為第1類的樣本正確的被判別到第1類的比率為99.15%；而{2-2}(特異度)的判別正確率較差，為94.08%。

表16、使用KSTAR分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	364(99.15%)	3(0.85%)
2 (有產生第二癌及有復發)	22(5.92%)	345(94.08%)

平均分類準確率：99.043%

由表17可知RandonizableFilteredClassifier的整體正確判別率為97.604%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為98.41%：即原始群體為第1類的樣本正確的被判別到第1類的比率為98.41%；而{2-2}(特異度)的判別正確率為59.35%。

表17、使用RandonizableFilteredClassifier分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	361(98.41%)	6(1.59%)
2 (有產生第二癌及有復發)	149(40.65%)	218(59.35%)

平均分類準確率：97.604%

由表18可知RandomTree的整體正確判別率為99.534%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為99.72%：即原始群體為第1類的樣本正確的被判別到第1類的比率為99.72%；而{2-2}(特異度)的判別正確率為92.88%。

表18、使用RandomTree分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	366(99.72%)	1(0.28%)
2 (有產生第二癌及有復發)	5(7.12%)	362(92.88%)

平均分類準確率：99.534%

從表 19 中，我們可以觀察到以有產生第二癌及有復發為目標變數，RandomTree 模式在{1-1}(敏感度)產生最高平均分類準確率，為 99.72%；而 IBK 在{2-2}(特異度)產生最高平均分類準確率，為 97.52%；在整體情況下，我們可以看到 RandomTree 模式優於 IBK、KSTAR 和 RandonizableFilteredClassifier 模式，這表明 RandomTree 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

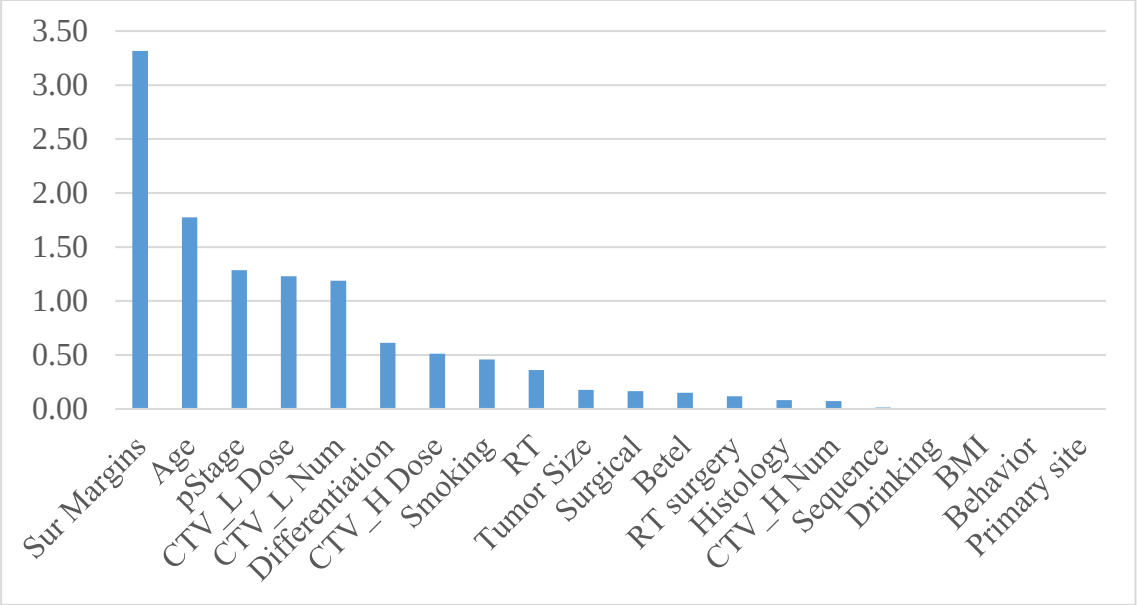
表19、IBK、KSTAR、RandonizableFilteredClassifier(RFC)和RandomTree(RT) 模式預測評估

	預測為其他，實際為其他（敏感度%）{1-1}				預測有產生第二癌及有復發，實際有產生第二癌及有復發（特異度%）{2-2}				整體平均預測準確率(%)			
模組	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT
1	99.50	99.25	98.17	99.83	96.55	96.15	52.17	94.12	99.43	99.18	97.30	99.67
2	99.41	99.25	98.33	99.33	96.43	92.59	63.64	92.86	99.35	99.10	97.71	99.18
3	99.58	99.33	98.33	99.50	96.67	92.86	56.00	93.33	99.51	99.18	97.47	99.35
4	99.50	99.08	98.50	99.83	96.55	95.83	72.73	88.89	99.43	99.02	98.04	99.51
5	99.50	99.16	97.91	99.83	100.0	92.31	37.50	94.12	99.51	99.02	96.73	99.67
6	99.50	99.08	98.91	100.0	100.0	95.83	60.00	94.44	99.51	99.02	97.79	99.84
7	99.25	98.83	98.25	99.75	96.15	90.91	54.17	96.88	99.18	98.69	97.38	99.67
8	99.33	99.08	98.42	99.83	96.30	95.83	65.22	88.89	99.26	99.02	97.79	99.51
9	99.50	99.16	98.82	99.58	96.55	92.31	62.50	85.29	99.43	99.02	97.87	99.18
10	99.50	99.25	98.50	99.75	100.0	96.15	69.57	100.0	99.51	99.18	97.96	99.75

接著使用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 分析後發現對於有產生第二癌且有復發的重要變數以手術邊緣(Surgical Margins)影響最大為 59.57%，原發部位(Primary Site)為最沒有影響力，如表 20 及圖六所示。

表20、Weka的InfoGain和GainRatio針對有產生第二癌且有復發的重要變數的重要性排序

Gain Ratio			Info Gain			平均重要變數之重要性排序		
重要性排序	百分比	重要變數	重要性排序	百分比	重要變數	重要性排序	百分比	重要變數
1	1.91	Sur Margins (手術邊緣)	1	1.41	Sur Margins (手術邊緣)	1	1.66	Sur Margins (手術邊緣)
2	0.97	Age (年齡)	2	0.81	Age (年齡)	2	0.89	Age (年齡)
3	0.80	CTV_L Dose (較低劑量)	3	0.64	pStage (病理期別)	3	0.65	pStage (病理期別)
4	0.78	CTV_L Num (較低次數)	4	0.43	CTV_L Dose (較低劑量)	4	0.62	CTV_L Dose (較低劑量)
5	0.65	pStage (病理期別)	5	0.41	CTV_L Num (較低次數)	5	0.60	CTV_L Num (較低次數)
6	0.29	CTV_H Dose (最高劑量)	6	0.38	Differentiation (分化)	6	0.31	Differentiation (分化)
7	0.24	Smoking (吸菸)	7	0.23	CTV_H Dose (最高劑量)	7	0.26	CTV_H Dose (最高劑量)
8	0.23	Differentiation (分化)	8	0.22	Smoking (吸菸)	8	0.23	Smoking (吸菸)
9	0.19	RT (放射)	9	0.17	RT (放射)	9	0.18	RT (放射)
10	0.14	Tumor Size (腫瘤大小)	10	0.08	Surgical (手術)	10	0.09	Tumor Size (腫瘤大小)
11	0.10	Betel (檳榔)	11	0.05	Betel (檳榔)	11	0.09	Surgical (手術)
12	0.10	RT surgery (手術前放射)	12	0.03	CTV_H Num (最高次數)	12	0.08	Betel (檳榔)
13	0.09	Surgical (手術)	13	0.03	Tumor Size (腫瘤大小)	13	0.06	RT surgery (手術前放射)
14	0.06	Histology (組織型態)	14	0.02	RT surgery (手術前放射)	14	0.04	Histology (組織型態)
15	0.04	CTV_H Num (最高次數)	15	0.02	Histology (組織型態)	15	0.04	CTV_H Num (最高次數)
16	0.01	Sequence (區域治療與全身性治療順序)	16	0.01	Sequence (區域治療與全身性治療順序)	16	0.01	Sequence (區域治療與全身性治療順序)
17	0.00	Drinking (喝酒)	17	0.00	Drinking (喝酒)	17	0.00	Drinking (喝酒)
18	0.00	BMI	18	0.00	BMI	18	0.00	BMI
19	0.00	Behavior Code (性態碼)	19	0.00	Behavior Code (性態碼)	19	0.00	Behavior Code (性態碼)
20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)



圖六、針對有產生第二癌且有復發的重要變數的重要性排序直方圖

5.4. 有第二癌但沒有復發預測分析結果

以有產生第二癌但沒有復發為目標變數，{1}代表：非有產生第二癌但沒有復發的其他 3 種情況；{2}則代表：有產生第二癌但沒有復發。因此{1-1}(敏感度)代表：原始的判定為其他 3 種情況，而經由模式判定後亦為其他 3 種情況；而{2-2}(特異度)則表示：原始判定為有產生第二癌但沒有復發，經由模式判定亦為有產生第二癌但沒有復發。

由表 21 知 IBK 的整體正確判別率為 99.182%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為 99.18%：即原始群體為第 1 類的樣本正確的被判別到第 1 類的比率為 99.18%。其中有 3 個原本群體為第 1 類的樣本，被錯分為第 2 類的群體中；而有 11 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表21、使用IBK分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌但沒有復發)
1(其他 3 種情況)	364(99.18%)	3(0.82%)
2 (有產生第二癌但沒有復發)	11(3.02%)	356(96.98%)

平均分類準確率：99.182%

由表22可知KSTAR的整體正確判別率為98.593%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為98.62%：即原始群體為第1類的樣本正確的被判別到第1類的比率為98.62%；而{2-2}(特異度)的判別正確率較差，為97.05%。

表22、使用KSTAR分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	362(98.62%)	5(1.38%)
2 (有產生第二癌但沒有復發)	11(2.95%)	356(97.05%)

平均分類準確率：98.593%

由表23可知RandonizableFilteredClassifier的整體正確判別率為95.168%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為96.91%：即原始群體為第1類的樣本正確的被判別到第1類的比率為96.91%；而{2-2}(特異度)的判別正確率為74.64%。

表23、使用RandonizableFilteredClassifier分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	356(96.91%)	11(3.09%)
2 (有產生第二癌但沒有復發)	93(25.36%)	274(74.64%)

平均分類準確率：95.168%

由表24可知RandomTree的整體正確判別率為99.395%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為99.25%：即原始群體為第1類的樣本正確的被判別到第1類的比率為99.25%；而{2-2}(特異度)的判別正確率為98.03%。

表24、使用RandomTree分類結果

	分類	
類別	1 (其他 3 種情況)	2 (有產生第二癌及有復發)
1(其他 3 種情況)	364(99.25%)	3(0.75%)
2 (有產生第二癌但沒有復發)	9(1.97%)	358(98.03%)

平均分類準確率：99.395%

從表 25 中，我們可以觀察到以有產生第二癌但沒有復發為目標變數，RandomTree 模式在{1-1}(敏感度)產生最高平均分類準確率，為 99.25%；而 RandomTree 在{2-2}(特異度)也產生最高平均分類準確率，為 98.03%；在整體情況下，我們可以看到 RandomTree 模式優於 IBK、KSTAR 和 RandonizableFilteredClassifier 模式，這表明 RandomTree 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

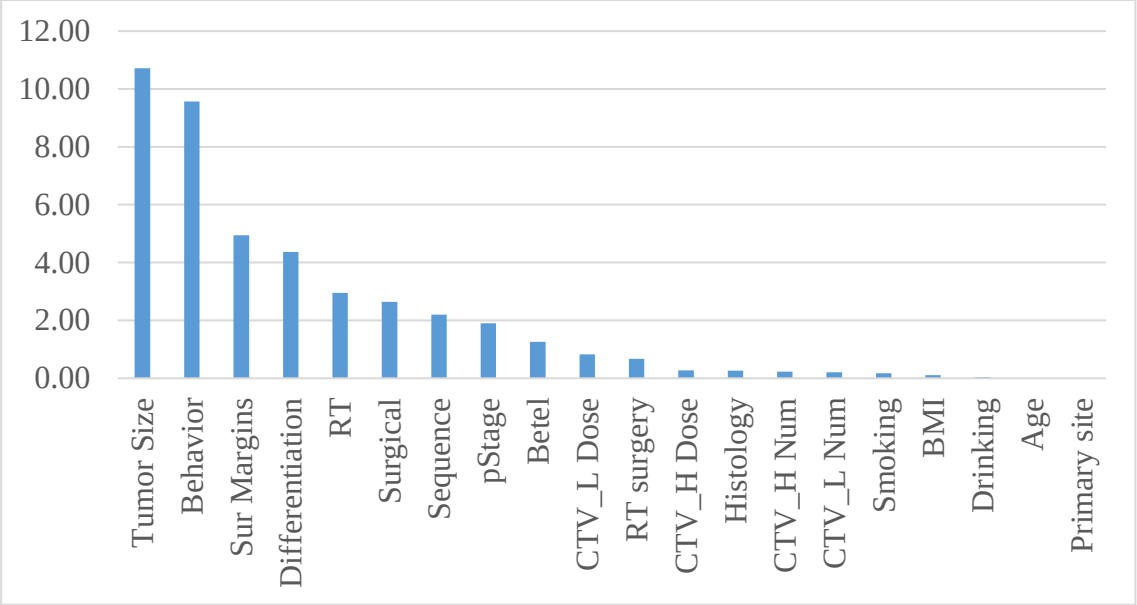
表25、IBK、KSTAR、RandonizableFilteredClassifier(RFC)和RandomTree(RT) 模式預測評估

	預測為其他，實際為其他（敏感度%）{1-1}				預測有產生第二癌但沒有復發，實際有產生第二癌但沒有復發（特異度%）{2-2}				整體平均預測準確率(%)			
模組	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT
1	99.55	98.74	96.30	99.37	96.72	97.32	71.93	97.48	99.26	98.61	94.03	99.18
2	96.69	97.35	98.17	95.76	99.46	98.83	52.17	99.10	99.18	98.69	97.30	98.77
3	99.64	99.10	96.77	99.91	95.20	95.76	81.31	96.83	99.18	98.77	95.42	99.59
4	99.46	98.65	96.85	99.55	96.69	96.43	79.28	97.52	99.18	98.45	95.26	99.35
5	99.37	98.56	97.02	100.0	96.67	97.27	76.92	99.19	99.10	98.45	95.09	99.92
6	99.28	98.65	97.76	99.64	96.64	97.30	91.59	97.54	99.02	98.53	97.22	99.43
7	99.28	98.56	96.92	99.55	96.64	96.40	74.17	99.16	99.02	98.36	94.69	99.51
8	99.55	98.83	96.20	99.36	96.72	96.49	68.64	95.08	99.26	98.61	93.54	98.94
9	99.55	98.74	96.94	99.64	97.52	97.32	80.18	100.0	99.35	98.61	95.42	99.67
10	99.46	99.01	96.12	99.73	97.50	97.39	70.18	98.36	99.26	98.86	93.70	99.59

接著使用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 分析後發現對於有產生第二癌但沒有復發的重要變數以腫瘤大小(Tumor Size) 影響最大為 10.72%，原發部位(Primary Site) 為最沒有影響力，如表 26 及圖七所示。

表26、Weka的InfoGain和GainRatio針對有產生第二癌但沒有復發的重要變數的重要性排序

Gain Ratio			Info Gain			平均重要變數之重要性排序		
重要性排序	百分比	重要變數	重要性排序	百分比	重要變數	重要性排序	百分比	重要變數
1	8.82	Tumor Size (腫瘤大小)	1	2.74	Differentiation (分化)	1	5.36	Tumor Size (腫瘤大小)
2	7.25	Behavior Code (性態碼)	2	2.32	Behavior Code (性態碼)	2	4.79	Behavior Code (性態碼)
3	2.84	Sur Margins (手術邊緣)	3	2.10	Sur Margins (手術邊緣)	3	2.48	Sur Margins (手術邊緣)
4	1.63	Differentiation (分化)	4	1.90	Tumor Size (腫瘤大小)	4	2.19	Differentiation (分化)
5	1.54	RT (放射)	5	1.41	RT (放射)	5	1.48	RT (放射)
6	1.39	Surgical (手術)	6	1.25	Surgical (手術)	6	1.32	Surgical (手術)
7	1.11	Sequence (區域治療與全身性治療順序)	7	1.09	Sequence (區域治療與全身性治療順序)	7	1.10	Sequence (區域治療與全身性治療順序)
8	0.96	pStage (病理期別)	8	0.95	pStage (病理期別)	8	0.95	pStage (病理期別)
9	0.86	Betel (檳榔)	9	0.40	Betel (檳榔)	9	0.63	Betel (檳榔)
10	0.54	RT surgery (手術前放射)	10	0.29	CTV_L Dose (較低劑量)	10	0.42	CTV_L Dose (較低劑量)
11	0.54	CTV_L Dose (較低劑量)	11	0.13	RT surgery (手術前放射)	11	0.34	RT surgery (手術前放射)
12	0.21	Histology (組織型態)	12	0.12	CTV_H Dose (最高劑量)	12	0.14	CTV_H Dose (最高劑量)
13	0.15	CTV_H Dose (最高劑量)	13	0.10	CTV_H Num (最高次數)	13	0.13	Histology (組織型態)
14	0.14	CTV_L Num (較低次數)	14	0.09	Smoking (吸菸)	14	0.12	CTV_H Num (最高次數)
15	0.13	CTV_H Num (最高次數)	15	0.07	CTV_L Num (較低次數)	15	0.11	CTV_L Num (較低次數)
16	0.09	Smoking (吸菸)	16	0.06	BMI	16	0.09	Smoking (吸菸)
17	0.06	BMI	17	0.05	Histology (組織型態)	17	0.06	BMI
18	0.02	Drinking (喝酒)	18	0.02	Drinking (喝酒)	18	0.02	Drinking (喝酒)
19	0.00	Age (年齡)	19	0.00	Age (年齡)	19	0.00	Age (年齡)
20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)



圖七、針對有產生第二癌但沒有復發的重要變數的重要性排序直方圖

5.5. 沒有第二癌但有復發預測分析結果

以沒有產生第二癌但有復發為目標變數，{1}代表：非沒有產生第二癌但有復發的其他 3 種情況；{2}則代表：沒有產生第二癌但有復發。因此{1-1}(敏感度)代表：原始的判定為其他 3 種情況，而經由模式判定後亦為其他 3 種情況；而{2-2}(特異度)則表示：原始判定為沒有產生第二癌但有復發，經由模式判定亦為沒有產生第二癌但有復發。

由表 27 知 IBK 的整體正確判別率為 99.052%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為 99.04%：即原始群體為第 1 類的樣本正確的被判別到第 1 類的比率為 99.04%。其中有 3 個原本群體為第 1 類的樣本，被錯分為第 2 類的群體中；而有 11 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表27、使用IBK分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌但有復發)
1(其他 3 種情況)	363(99.04%)	4(0.96%)
2 (沒有產生第二癌但有復發)	6(1.63%)	361(98.37%)

平均分類準確率：99.052%

由表28可知KSTAR的整體正確判別率為98.700%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為98.69%：即原始群體為第1類的樣本正確的被判別到第1類的比率為98.69%；而{2-2}(特異度)的判別正確率較差，為97.89%。

表28、使用KSTAR分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌但有復發)
1(其他 3 種情況)	362(98.69%)	5(1.31%)
2 (沒有產生第二癌但有復發)	8(2.11%)	359(97.89%)

平均分類準確率：98.700%

由表29可知RandonizableFilteredClassifier的整體正確判別率為91.813%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為90.96%：即原始群體為第1類的樣本正確的被判別到第1類的比率為90.96%；而{2-2}(特異度)的判別正確率為89.11%。

表29、使用RandonizableFilteredClassifier分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌但有復發)
1(其他 3 種情況)	334(90.96%)	33(9.04%)
2 (沒有產生第二癌但有復發)	40(10.89%)	327(89.11%)

平均分類準確率：91.813%

由表30可知RandomTree的整體正確判別率為98.839%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為98.78%：即原始群體為第1類的樣本正確的被判別到第1類的比率為98.78%；而{2-2}(特異度)的判別正確率為98.11%。

表30、使用RandomTree分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌但有復發)
1(其他 3 種情況)	363(98.78%)	4(0.75%)
2 (沒有產生第二癌但有復發)	7(1.97%)	360(98.11%)

平均分類準確率：98.839%

從表 31 中，我們可以觀察到以沒有產生第二癌但有復發為目標變數，IBK 模式在{1-1}(敏感度)產生最高平均分類準確率，為 99.04%；而 IBK 在{2-2}(特異度)也產生最高平均分類準確率，為98.37%；在整體情況下，我們可以看到 IBK 模式優於 KSTAR、RandonizableFilteredClassifier 和 RandomTree 模式，這表明 IBK 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

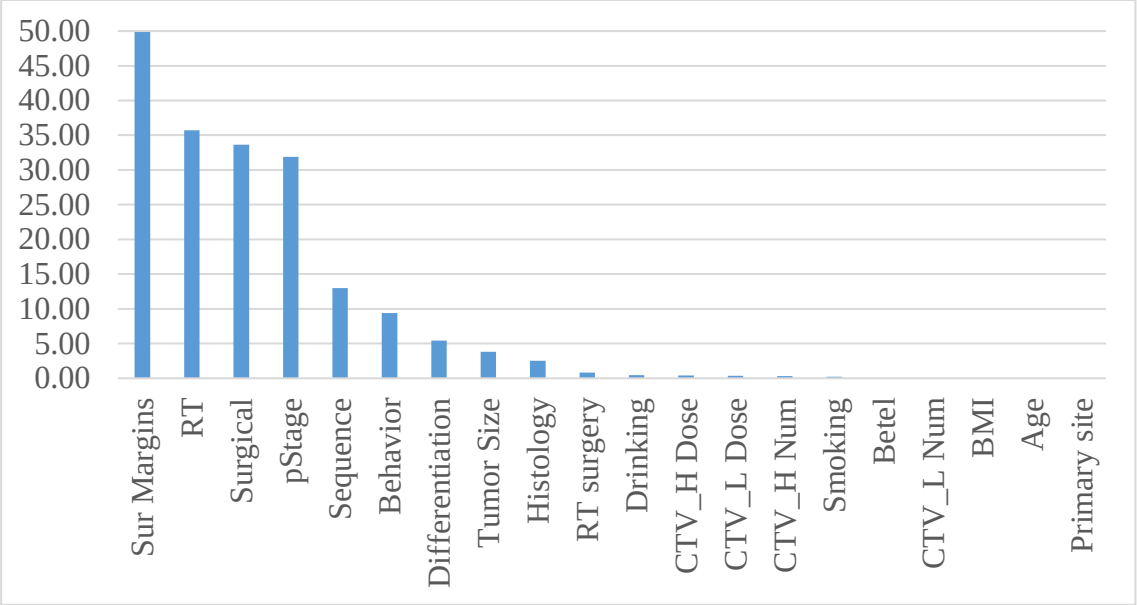
表31、IBK、KSTAR、RandonizableFilteredClassifier(RFC)和RandomTree(RT) 模式預測評估

	預測為其他，實際為其他（敏感度%）{1-1}				預測沒有產生第二癌但有復發，實際沒有產生第二癌但有復發（特異度%）{2-2}				整體平均預測準確率(%)			
模組	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT
1	98.23	97.35	84.98	97.03	99.66	99.43	94.04	98.98	99.26	98.86	91.58	98.45
2	99.55	99.43	91.30	98.99	97.94	97.35	81.59	98.49	99.10	98.86	88.80	98.86
3	99.32	99.10	95.84	99.55	97.63	96.76	89.79	96.79	98.86	98.45	94.19	98.77
4	99.43	99.21	94.44	99.89	97.64	97.05	83.92	99.11	98.94	98.61	91.50	99.67
5	97.94	97.03	83.09	97.94	99.55	98.98	93.68	99.55	99.10	98.45	90.76	99.10
6	99.32	99.21	92.77	98.87	97.92	97.34	93.15	97.31	98.94	98.69	92.95	98.45
7	99.66	99.32	94.03	99.43	97.94	97.92	84.48	96.50	99.18	98.94	91.41	98.61
8	99.32	99.10	94.18	99.32	97.63	97.33	86.32	97.63	98.86	98.61	92.07	98.86
9	98.22	97.91	83.69	97.35	99.55	99.10	93.39	99.32	99.18	98.77	90.76	98.77
10	99.44	99.21	95.33	99.43	98.22	97.63	90.74	97.35	99.10	98.77	94.11	98.86

再來使用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 分析後發現對於沒有產生第二癌但有復發的重要變數以手術邊緣(Surgical Margins) 影響最大為 49.84%，原發部位(Primary Site) 為最沒有影響力，如表 32 及圖八所示。

表32、Weka的InfoGain和GainRatio針對沒有產生第二癌但有復發的重要變數的重要性排序

Gain Ratio			Info Gain			平均重要變數之重要性排序		
重要性排序	百分比	重要變數	重要性排序	百分比	重要變數	重要性排序	百分比	重要變數
1	28.64	Sur Margins (手術邊緣)	1	21.20	Sur Margins (手術邊緣)	1	24.92	Sur Margins (手術邊緣)
2	18.64	RT (放射)	2	17.07	RT (放射)	2	17.86	RT (放射)
3	17.66	Surgical (手術)	3	15.95	Surgical (手術)	3	16.81	Surgical (手術)
4	16.03	pStage (病理期別)	4	15.86	pStage (病理期別)	4	15.95	pStage (病理期別)
5	7.11	Behavior Code (性態碼)	5	6.41	Sequence (區域治療與全身性治療順序)	5	6.49	Sequence (區域治療與全身性治療順序)
6	6.56	Sequence (區域治療與全身性治療順序)	6	3.39	Differentiation (分化)	6	4.69	Behavior Code (性態碼)
7	3.14	Tumor Size (腫瘤大小)	7	2.27	Behavior Code (性態碼)	7	2.71	Differentiation (分化)
8	2.02	Differentiation (分化)	8	0.68	Tumor Size (腫瘤大小)	8	1.91	Tumor Size (腫瘤大小)
9	2.00	Histology (組織型態)	9	0.52	Histology (組織型態)	9	1.26	Histology (組織型態)
10	0.68	RT surgery (手術前放射)	10	0.26	Drinking (喝酒)	10	0.42	RT surgery (手術前放射)
11	0.30	CTV_L Dose (較低劑量)	11	0.18	CTV_H Dose (最高劑量)	11	0.22	Drinking (喝酒)
12	0.23	CTV_H Dose (最高劑量)	12	0.16	RT surgery (手術前放射)	12	0.21	CTV_H Dose (最高劑量)
13	0.18	Drinking (喝酒)	13	0.15	CTV_H Num (最高次數)	13	0.19	CTV_L Dose (較低劑量)
14	0.14	CTV_H Num (最高次數)	14	0.10	Smoking (吸菸)	14	0.15	CTV_H Num (最高次數)
15	0.11	Smoking (吸菸)	15	0.08	CTV_L Dose (較低劑量)	15	0.11	Smoking (吸菸)
16	0.03	Betel (檳榔)	16	0.01	Betel (檳榔)	16	0.02	Betel (檳榔)
17	0.02	CTV_L Num (較低次數)	17	0.01	CTV_L Num (較低次數)	17	0.02	CTV_L Num (較低次數)
18	0.01	BMI	18	0.01	BMI	18	0.01	BMI
19	0.00	Age (年齡)	19	0.00	Age (年齡)	19	0.00	Age (年齡)
20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)



圖八、針對沒有產生第二癌但有復發的重要變數的重要性排序直方圖

5.6. 沒有第二癌也沒有復發預測分析結果

以沒有產生第二癌也沒有復發為目標變數，{1}代表：非沒有產生第二癌也沒有復發的其他 3 種情況；{2}則代表：沒有產生第二癌也沒有復發。因此{1-1}(敏感度)代表：原始的判定為其他 3 種情況，而經由模式判定後亦為其他 3 種情況；而{2-2}(特異度)則表示：原始判定為沒有產生第二癌也沒有復發，經由模式判定亦為沒有產生第二癌也沒有復發。

由表 33 知 IBK 的整體正確判別率為 98.970%，而個別的判別正確率以{2-2}(特異度)的比率最高，為 99.18%：即原始群體為第 2 類的樣本正確的被判別到第 2 類的比率為 99.18%。其中有 3 個原本群體為第 1 類的樣本，被錯分為第 2 類的群體中；而有 11 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表33、使用IBK分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌也沒有復發)
1(其他 3 種情況)	363(98.90%)	4(1.1%)
2 (沒有產生第二癌也沒有復發)	3(082%)	364(99.18%)

平均分類準確率：98.970%

由表34可知KSTAR的整體正確判別率為98.373%，而個別的判別正確率以{2-2}(特異度)的比率最高，為98.67%：即原始群體為第2類的樣本正確的被判別到第2類的比率為98.67%；而{1-1}(敏感度)的判別正確率較差，為98.31%。

表34、使用KSTAR分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌也沒有復發)
1(其他 3 種情況)	361(98.31%)	6(1.69%)
2 (沒有產生第二癌也沒有復發)	5(1.33%)	362(98.67%)

平均分類準確率：98.373%

由表35可知RandonizableFilteredClassifier的整體正確判別率為90.890%，而個別的判別正確率以{2-2}(特異度)的比率最高，為91.31%：即原始群體為第2類的樣本正確的被判別到第2類的比率為91.31；而{1-1}(敏感度)的判別正確率為90.60%。

表35、使用RandonizableFilteredClassifier分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌也沒有復發)
1(其他 3 種情況)	333(90.60%)	34(9.04%)
2 (沒有產生第二癌也沒有復發)	33(10.89%)	335(91.31%)

平均分類準確率：90.890%

由表36可知RandomTree的整體正確判別率為99.256%，而個別的判別正確率以{1-1}(敏感度)的比率最高，為99.32%：即原始群體為第1類的樣本正確的被判別到第1類的比率為99.32%；而{2-2}(特異度)的判別正確率為99.18%。

表36、使用RandomTree分類結果

	分類	
類別	1 (其他 3 種情況)	2 (沒有產生第二癌也沒有復發)
1(其他 3 種情況)	365(99.32%)	2(0.68%)
2 (沒有產生第二癌也沒有復發)	3(0.82%)	364(99.18%)

平均分類準確率：99.256%

從表 37 中，我們可以觀察到以沒有產生第二癌也沒有復發為目標變數，RandomTree 模式在{1-1}(敏感度)產生最高平均分類準確率，為 99.32%；而 RandomTree 在{2-2}(特異度)也產生最高平均分類準確率，為 99.18%；在整體情況下，我們可以看到 RandomTree 模式優於 IBK、KSTAR 和 RandonizableFilteredClassifier 模式，這表明 RandomTree 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

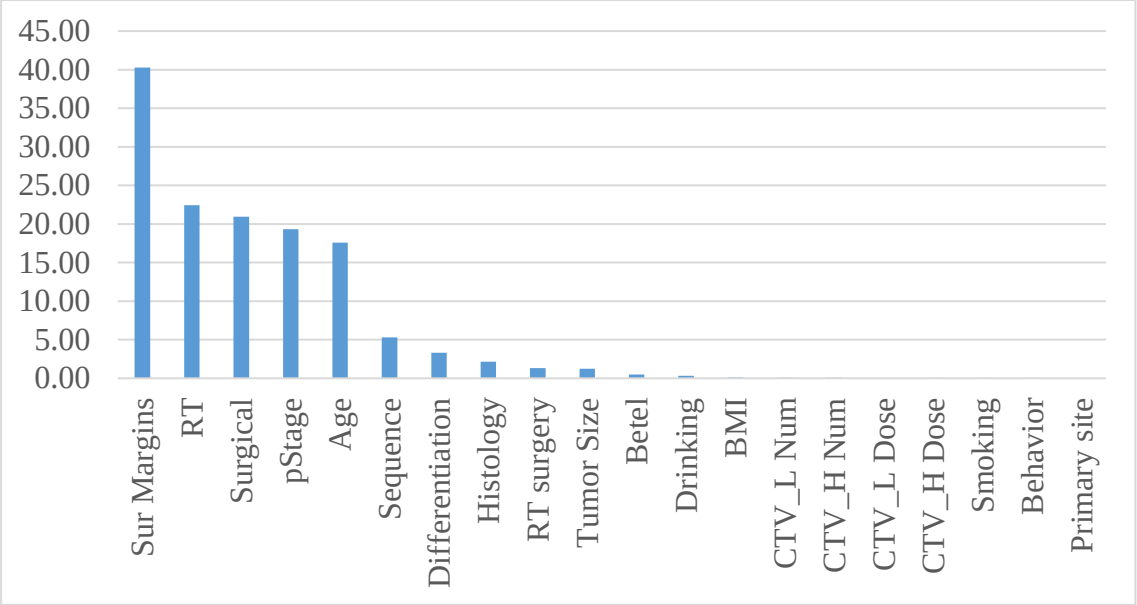
表37、IBK、KSTAR、RandonizableFilteredClassifier(RFC)和RandomTree(RT) 模式預測評估

	預測為其他，實際為其他（敏感度%）{1-1}				預測沒有產生第二癌也沒有復發，實際沒有產生第二癌也沒有復發（特異度%）{2-2}				整體平均預測準確率(%)			
模組	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT	IBK	KSTAR	RFC	RT
1	99.38	98.97	88.82	99.18	98.78	98.24	90.39	98.64	99.02	98.53	89.78	98.86
2	99.39	98.96	90.95	98.77	99.05	97.97	93.08	98.64	99.18	98.36	92.23	98.69
3	98.64	97.98	88.28	99.73	99.38	99.17	85.81	99.59	98.94	98.45	87.33	99.67
4	98.51	97.84	93.90	99.86	99.38	99.17	92.37	99.60	98.86	98.36	93.30	99.75
5	99.59	99.38	91.52	98.59	98.91	97.98	89.29	99.31	99.18	98.53	90.11	99.02
6	98.64	97.71	90.46	99.86	99.38	98.96	89.96	99.19	98.94	98.20	90.27	99.59
7	98.64	97.58	90.19	98.91	99.38	98.96	100.0	99.18	98.94	98.12	93.69	99.02
8	98.51	98.10	90.51	99.18	99.38	98.76	90.73	99.19	98.86	98.36	90.60	99.18
9	99.18	98.97	90.68	99.59	98.78	98.37	91.34	98.91	98.94	98.61	91.09	99.18
10	98.51	97.58	90.72	99.59	99.38	99.16	90.19	99.59	98.86	98.20	90.52	99.59

最後使用 Weka 的 InfoGainAttributeEval 和 GainRatioAttributeEval 分析後發現對於沒有產生第二癌且沒有復發的重要變數以手術邊緣(Surgical Margins) 影響最大為 49.84%，原發部位(Primary Site)為最沒有影響力，如表 38 及圖九所示。

表38、Weka的InfoGain和GainRatio針對沒有產生第二癌且沒有復發的重要變數的重要性排序

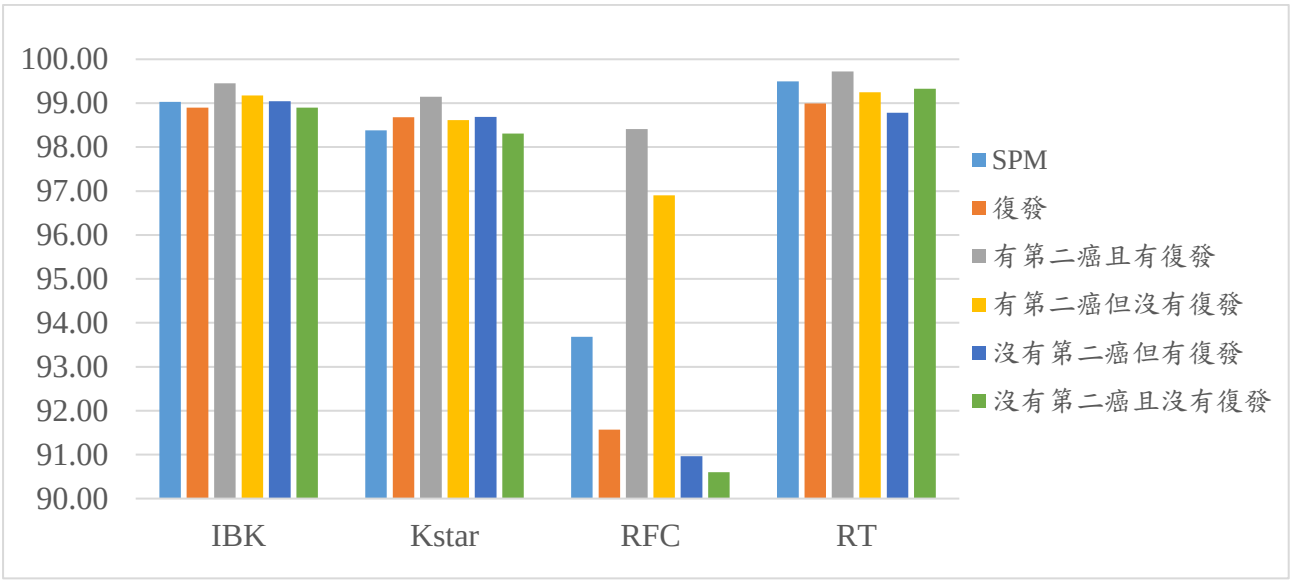
Gain Ratio			Info Gain			平均重要變數之重要性排序		
重要性排序	百分比	重要變數	重要性排序	百分比	重要變數	重要性排序	百分比	重要變數
1	23.13	Sur Margins (手術邊緣)	1	17.12	Sur Margins (手術邊緣)	1	20.13	Sur Margins (手術邊緣)
2	16.30	Age (年齡)	2	10.73	RT (放射)	2	11.23	RT (放射)
3	11.72	RT (放射)	3	9.95	Surgical (手術)	3	10.48	Surgical (手術)
4	11.01	Surgical (手術)	4	9.61	pStage (病理期別)	4	9.67	pStage (病理期別)
5	9.72	pStage (病理期別)	5	2.62	Sequence (區域治療與全身性治療順序)	5	8.80	Age (年齡)
6	2.68	Sequence (區域治療與全身性治療順序)	6	2.07	Differentiation (分化)	6	2.65	Sequence (區域治療與全身性治療順序)
7	1.71	Histology (組織型態)	7	1.30	Age (年齡)	7	1.65	Differentiation (分化)
8	1.23	Differentiation (分化)	8	0.44	Histology (組織型態)	8	1.08	Histology (組織型態)
9	1.08	RT surgery (手術前放射)	9	0.25	RT surgery (手術前放射)	9	0.67	RT surgery (手術前放射)
10	1.01	Tumor Size (腫瘤大小)	10	0.22	Tumor Size (腫瘤大小)	10	0.62	Tumor Size (腫瘤大小)
11	0.32	Betel (檳榔)	11	0.15	Betel (檳榔)	11	0.24	Betel (檳榔)
12	0.17	Drinking (喝酒)	12	0.15	Drinking (喝酒)	12	0.16	Drinking (喝酒)
13	0.06	CTV_L Num (較低次數)	13	0.05	BMI	13	0.05	BMI
14	0.05	BMI	14	0.03	CTV_L Num (較低次數)	14	0.05	CTV_L Num (較低次數)
15	0.04	CTV_H Num (最高次數)	15	0.03	CTV_H Num (最高次數)	15	0.04	CTV_H Num (最高次數)
16	0.03	CTV_L Dose (較低劑量)	16	0.02	CTV_L Dose (較低劑量)	16	0.03	CTV_L Dose (較低劑量)
17	0.01	CTV_H Dose (最高劑量)	17	0.01	CTV_H Dose (最高劑量)	17	0.01	CTV_H Dose (最高劑量)
18	0.00	Smoking (吸菸)	18	0.00	Smoking (吸菸)	18	0.01	Smoking (吸菸)
19	0.00	Behavior Code (性態碼)	19	0.00	Behavior Code (性態碼)	19	0.00	Behavior Code (性態碼)
20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)	20	0.00	Primary Site (原發部位)



圖九、針對沒有產生第二癌且沒有復發的重要變數的重要性排序直方圖

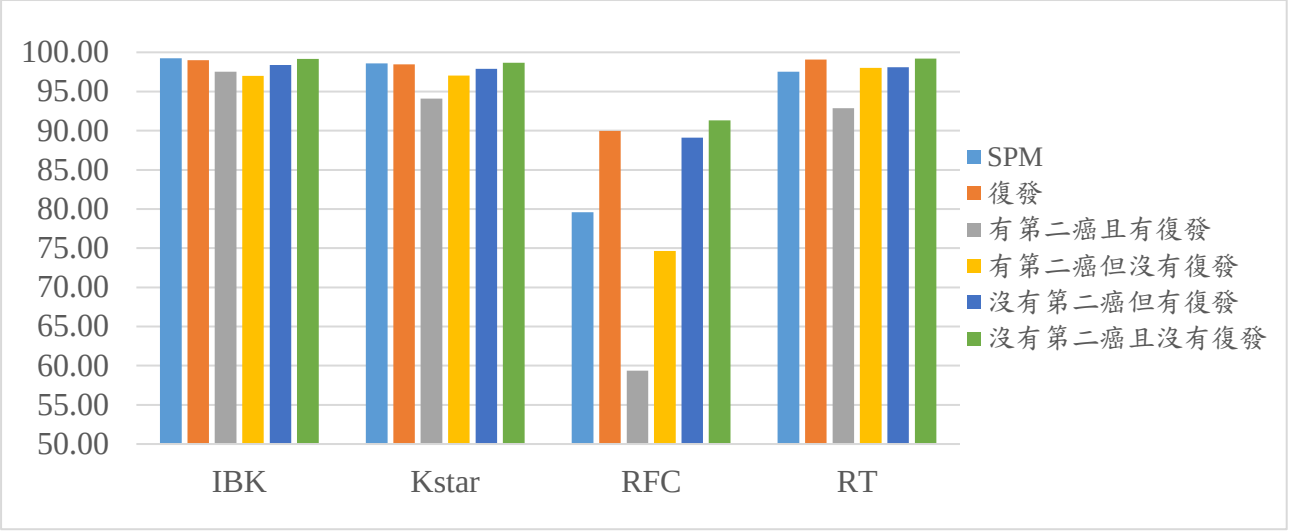
5.7. 4種方法的分析結果

第三階段，我們使用 IBK、KSTAR、RandonizableFilteredClassifier 和 RandomTree 方法對數據集中共包含 19 個預測變數以及 6 個目標變數驗證其敏感度、特異度和方法之準確率。經過 4 個方法之比較後發現，RandomTree 方法針對 6 個目標變數之敏感度皆在 98%以上，如圖九所示。



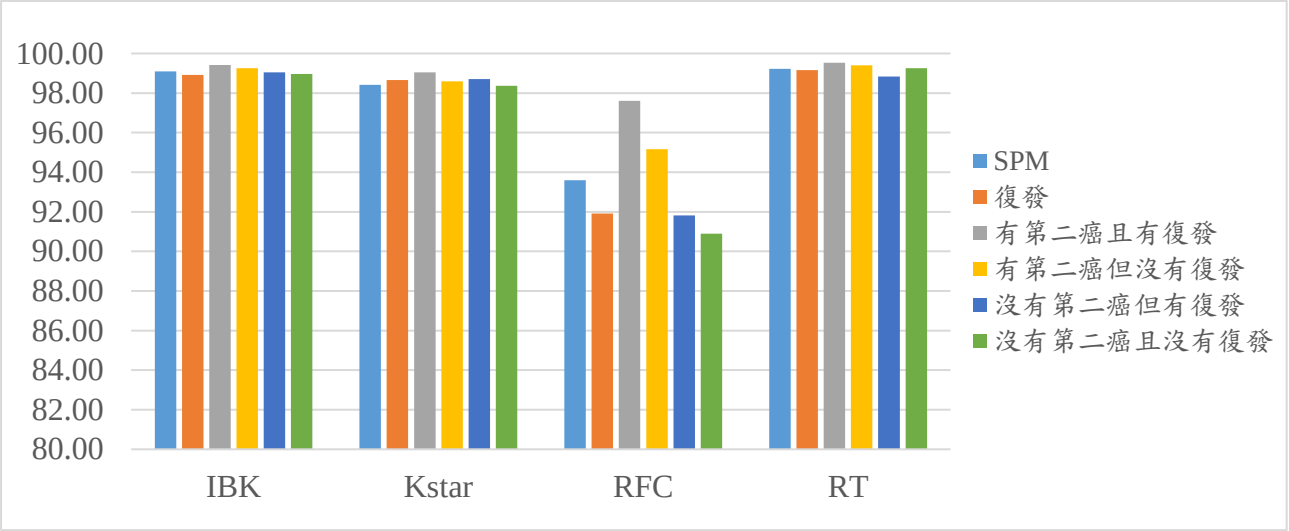
圖九、4種方法之敏感度的比較

針對 4 種方法之特異度的比較，IBK 方法有較高百分比，對於 6 個目標變數之百分比皆為 96%以上(最高特異度為 99.24%)，其次依序為 RandomTree (最高特異度為 99.07%)、KSTAR (最高特異度為 98.67%)、RandonizableFilteredClassifier (最高特異度為 91.32%)，如圖十所示。



圖十、4種方法之特異度的比較

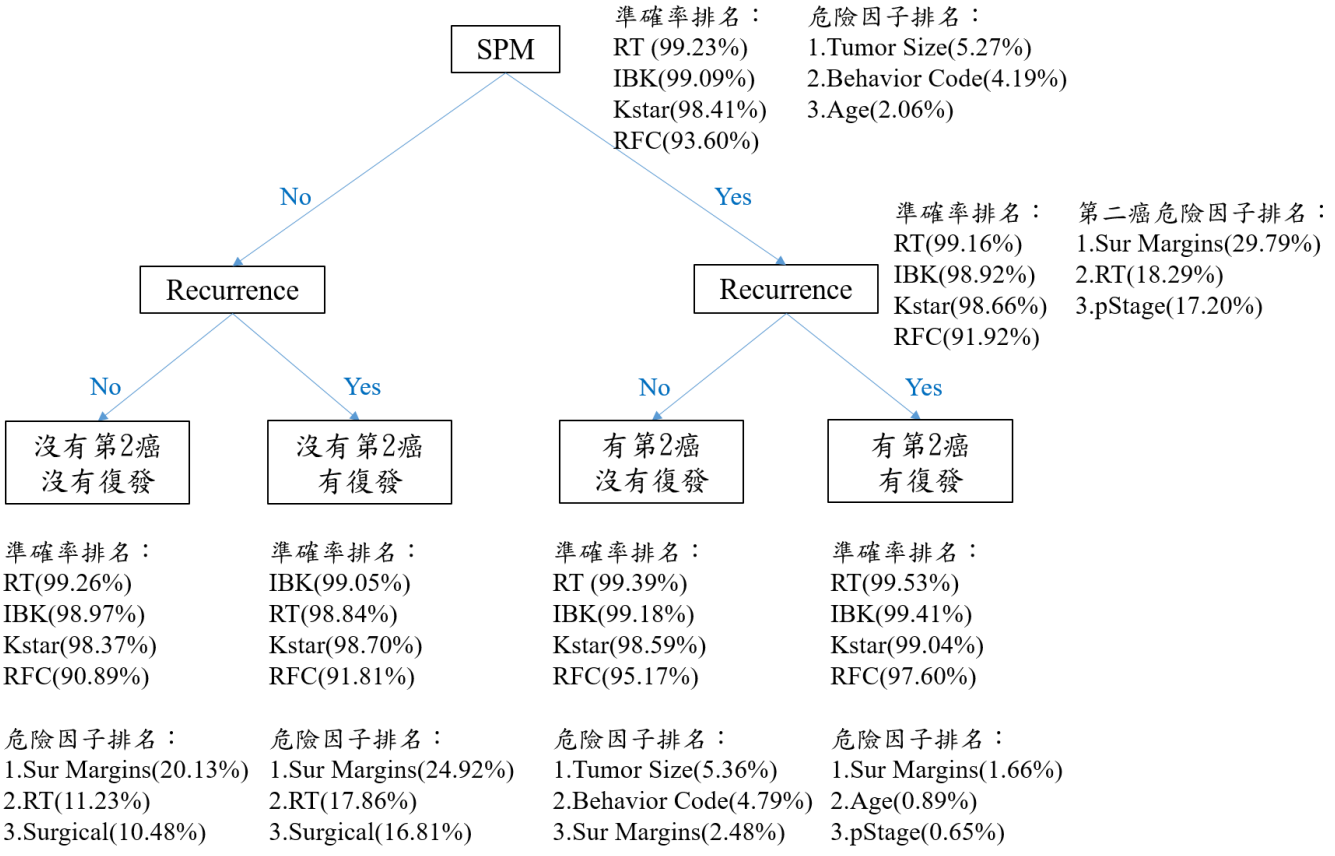
針對 4 種方法之平均準確率的比較，RandomTree 方法有較高百分比，對於 6 個目標變數之百分比皆為 98%以上(最高平均準確率為 99.53%)，其次依序為 IBK(最高平均準確率為 99.41%)、KSTAR(最高平均準確率為 99.04%)、RandonizableFilteredClassifier(最高平均準確率為 97.60%)，如圖十一所示，統整之總表如表 39 所示，分層結果圖如圖十二所示。



圖十一、4種方法之平均準確率的比較

表39、4種方法針對6個目標變數的結果的比較

目標變數	結果 / 方法	敏感度	特異度	準確率	Fallout	F1_score	MCC
SPM	IBK	99.03	**99.24**	99.09	5.75	99.17	95.90
	KStar	98.38	98.60	98.41	10.21	98.64	92.76
	RFC	93.68	79.61	93.60	27.35	94.24	70.29
	RT	**99.49**	97.53	**99.23**	3.44	99.50	96.55
復發	IBK	98.90	99.01	98.92	2.04	98.95	97.44
	KStar	98.68	98.45	98.66	2.13	98.65	96.82
	RFC	91.57	89.97	91.92	11.21	91.27	80.67
	RT	**98.99**	**99.07**	**99.16**	1.01	98.98	98.01
有第二癌且有復發	IBK	99.46	**97.52**	99.41	19.12	99.70	88.52
	KStar	99.15	94.08	99.04	30.00	99.51	80.67
	RFC	98.41	59.35	97.60	55.88	98.77	49.82
	RT	**99.72**	92.88	**99.53**	9.71	99.76	91.27
有第二癌但沒有復發	IBK	99.18	96.98	99.18	4.43	99.18	95.45
	KStar	98.62	97.05	98.59	10.11	98.62	92.03
	RFC	96.91	74.64	95.17	32.73	97.34	68.16
	RT	**99.25**	**98.03**	**99.39**	2.97	99.11	96.63
沒有第二癌但有復發	IBK	**99.04**	**98.37**	**99.05**	1.26	99.07	97.63
	KStar	98.69	97.89	98.70	1.72	98.66	96.74
	RFC	90.97	89.11	91.81	11.93	91.14	79.73
	RT	98.78	98.11	98.84	1.48	98.78	97.10
沒有第二癌且沒有復發	IBK	98.90	99.18	98.97	1.44	98.99	97.86
	KStar	98.31	98.67	98.37	2.23	98.40	96.62
	RFC	90.60	91.32	90.89	10.99	90.85	81.06
	RT	**99.33**	**99.18**	**99.26**	0.69	99.20	98.45



圖十二、分層結果圖

(六) 結論

為了獲得更佳的大腸直腸癌復發之重要因子，本專題研究使用多種資料探勘方法找出復發及產生第二癌的危險因素。研究結果支持手術邊緣(Surgical Margins) 和腫瘤大小(Tumor Size) 是重要的預測復發影響因子。進一步比較IBK、KSTAR、RandonizableFilteredClassifier和RandomTree方法在大腸直腸癌的預測準確度。本研究發現RandomTree方法針對6個目標變數之預測敏感度為最高，百分比皆在98%以上，IBK方法對於6個目標變數之預測特異度，百分比皆為96%以上，而RandomTree方法對於6個目標變數之平均準確率為最高，百分比皆為98%以上本研究結果證實針對大腸直腸癌症病患的復發性及產生第二癌之預測，資料分析結果證實可以手術邊緣和腫瘤大小作為基礎，再經由本研究架構的分層樣式提供臨床醫師輔助治療。

(七)文獻參考

Leung W. H. and Liu C. K. (2014) Chemotherapy and Targeted Therapy in Colorectal Cancer: The Current Status, Journal of Cancer Research and Practice, Vol. 30, no. 1, pp. 11-20.

Li F. G., Wang Z. P., Hu G. and Li H. (2011) Current status of SNPs interaction in genome-wide association study, Yi Chuan, Vol. 33, no. 9, pp. 901-10.

Ong L. S., Shepherd B., Tong L. C., Seow-Choen F., Ho Y. H., Tang C. L. and Tan K. (1997) The colorectal cancer recurrence support (CARES) system, Artificial Intelligence in Medicine, Vol. 11, no. 3, pp. 175-188.

Vani G., Savitha R. and Sundararajan N. (2010) Classification of Abnormalities in Digitized Mammograms using Extreme Learning Machine, Automation, Robotics and Vision Singapore.

Walker A. S., Johnson E. K., Maykel J. A., Stojadinovic A., Nissan A., Brucher B. and Steele S. R. (2014) Future Directions for the Early Detection of Colorectal Cancer Recurrence, Journal of Cancer, Vol. 5, no. 4, pp. 272-280.

International Agency for Research on Cancer (IARC)(2016)，取自http://globocan.iarc.fr/Pages/Map.aspx.

Health Promotion Administration Miistry of Health and Welfare，取自 http://www.hpa.gov.tw/EngPages/Index.aspx.

張華偉，王明文 & 甘麗新。(2006)。基於隨機森林的文本分類模型研究. 山東大學學報: 理學版, 41(3), pp.139-143.