FDA TermProject

Project:登革熱

Team: shallow learning

Members:薛勝謙、簡瑜成

目錄

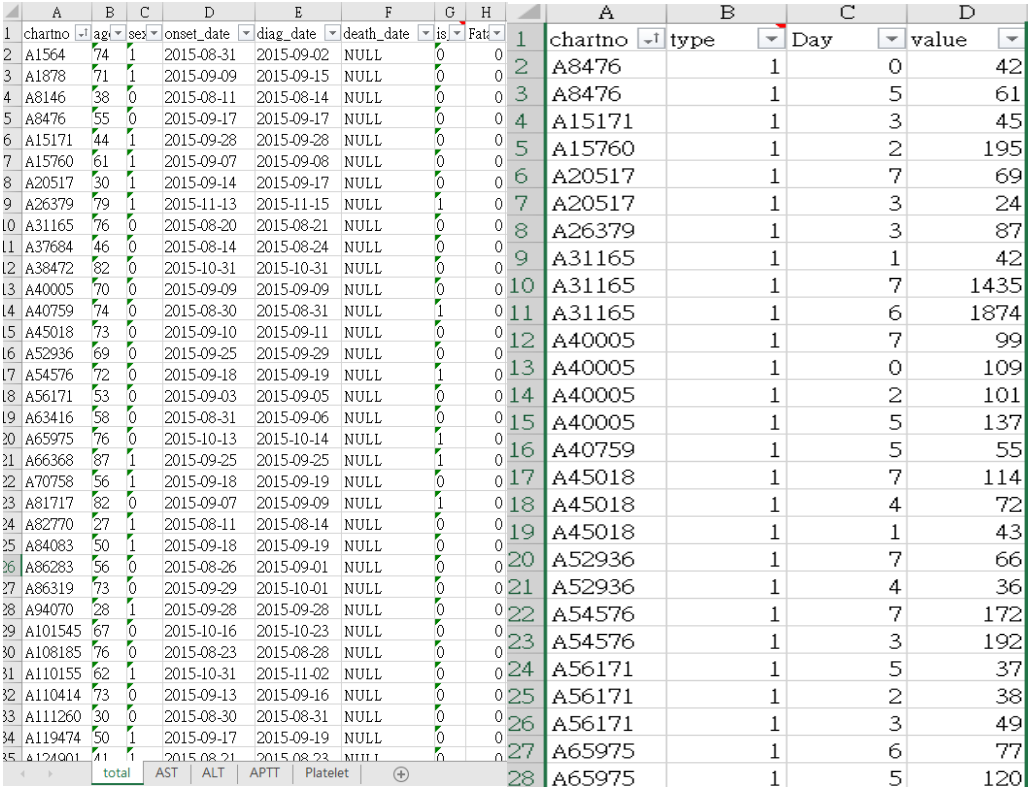
分析主題

此次拿到的資料為登革熱的病患病歷資料，其中包含年齡、性別等基本資料，以及就診日期、各項醫療指數(肝功能指數與血小板指數)；最後則是病患死亡與否。因次我們決定最終目標為根據病歷資料，及早預測病患的死亡可能，達到盡早預防的效果，並提供給醫師做為參考憑據。

因此將 Fatal 設定為 label，將根據其餘數據特徵做二元分類，預測 Fatal 為 0 或 1，代表死亡或存活。首先將採用決策樹模型，以對學習預測效果有綜觀及初步的成果檢視，並希望以較接近人類做決策的方式，提供給醫護人員參考標準；之後將用 SVM 模型來互相比較正確率。

資料處理、分析與模型建立

- 原始資料:



	A	B	C	D	E	F	G	H
1	chartno	ag	sex	onset_date	diag_date	death_date	is	Fat
2	A1564	74	1	2015-08-31	2015-09-02	NULL	0	0
3	A1878	71	1	2015-09-09	2015-09-15	NULL	0	0
4	A8146	38	0	2015-08-11	2015-08-14	NULL	0	0
5	A8476	55	0	2015-09-17	2015-09-17	NULL	0	0
6	A15171	44	1	2015-09-28	2015-09-28	NULL	0	0
7	A15760	61	1	2015-09-07	2015-09-08	NULL	0	0
8	A20517	30	1	2015-09-14	2015-09-17	NULL	0	0
9	A26379	79	1	2015-11-13	2015-11-15	NULL	1	0
10	A31165	76	0	2015-08-20	2015-08-21	NULL	0	0
11	A37684	46	0	2015-08-14	2015-08-24	NULL	0	0
12	A38472	82	0	2015-10-31	2015-10-31	NULL	0	0
13	A40005	70	0	2015-09-09	2015-09-09	NULL	0	0
14	A40759	74	0	2015-08-30	2015-08-31	NULL	1	0
15	A45018	73	0	2015-09-10	2015-09-11	NULL	0	0
16	A52936	69	0	2015-09-25	2015-09-29	NULL	0	0
17	A54576	72	0	2015-09-18	2015-09-19	NULL	1	0
18	A56171	53	0	2015-09-03	2015-09-05	NULL	0	0
19	A63416	58	0	2015-08-31	2015-09-06	NULL	0	0
20	A65975	76	0	2015-10-13	2015-10-14	NULL	1	0
21	A66368	87	1	2015-09-25	2015-09-25	NULL	1	0
22	A70758	56	1	2015-09-18	2015-09-19	NULL	0	0
23	A81717	82	0	2015-09-07	2015-09-09	NULL	1	0
24	A82770	27	1	2015-08-11	2015-08-14	NULL	0	0
25	A84083	50	1	2015-09-18	2015-09-19	NULL	0	0
26	A86283	56	0	2015-08-26	2015-09-01	NULL	0	0
27	A86319	73	0	2015-09-29	2015-10-01	NULL	0	0
28	A94070	28	1	2015-09-28	2015-09-28	NULL	0	0
29	A101545	67	0	2015-10-16	2015-10-23	NULL	0	0
30	A108185	76	0	2015-08-23	2015-08-28	NULL	0	0
31	A110155	62	1	2015-10-31	2015-11-02	NULL	0	0
32	A110414	73	0	2015-09-13	2015-09-16	NULL	0	0
33	A111260	30	0	2015-08-30	2015-08-31	NULL	0	0
34	A119474	50	1	2015-09-17	2015-09-19	NULL	0	0
35	A124901	41	1	2015-08-21	2015-08-23	NULL	0	0

total | AST | ALT | APTT | Platelet

	A	B	C	D
1	chartno	type	Day	value
2	A8476	1	0	42
3	A8476	1	5	61
4	A15171	1	3	45
5	A15760	1	2	195
6	A20517	1	7	69
7	A20517	1	3	24
8	A26379	1	3	87
9	A31165	1	1	42
10	A31165	1	7	1435
11	A31165	1	6	1874
12	A40005	1	7	99
13	A40005	1	0	109
14	A40005	1	2	101
15	A40005	1	5	137
16	A40759	1	5	55
17	A45018	1	7	114
18	A45018	1	4	72
19	A45018	1	1	43
20	A52936	1	7	66
21	A52936	1	4	36
22	A54576	1	7	172
23	A54576	1	3	192
24	A56171	1	5	37
25	A56171	1	2	38
26	A56171	1	3	49
27	A65975	1	6	77
28	A65975	1	5	120

原始資料具有五個工作表，total 包含病歷號及年齡等基本資料，

其餘工作表皆為病歷號與不同醫療指數、死亡與否、檢測日期。

接著程式處理方面，程式架構請參閱 readme，以下僅作概念闡述。

- 資料前處理:

因為各病歷號所量測的醫療指數並不完整，故各工作表的資料量皆不相同，因此我們先將各病歷號與全部的醫療指數整理到同一張表，缺漏值則採用隨機森林回歸方法，來較有規則的填補。並將日期整理成發病日減去確診日的天數。

chartno	age	sex	is_hospitalization	Fatal	AST_value	ALT_value	APTT_value	Platelet_value	diag-onset
A10015442	36	1	0	0	39.0	25.0	36.600000	98.301667	2.0
A10017629	35	1	0	0	58.0	42.0	33.200000	83.650000	3.0
A10030438	75	0	1	0	45.0	10.0	40.500000	51.000000	0.0
A10031096	65	1	0	0	81.0	41.0	39.818000	120.000000	0.0
A10034524	58	0	0	0	27.0	13.0	35.936833	93.000000	2.0

結果如上圖。

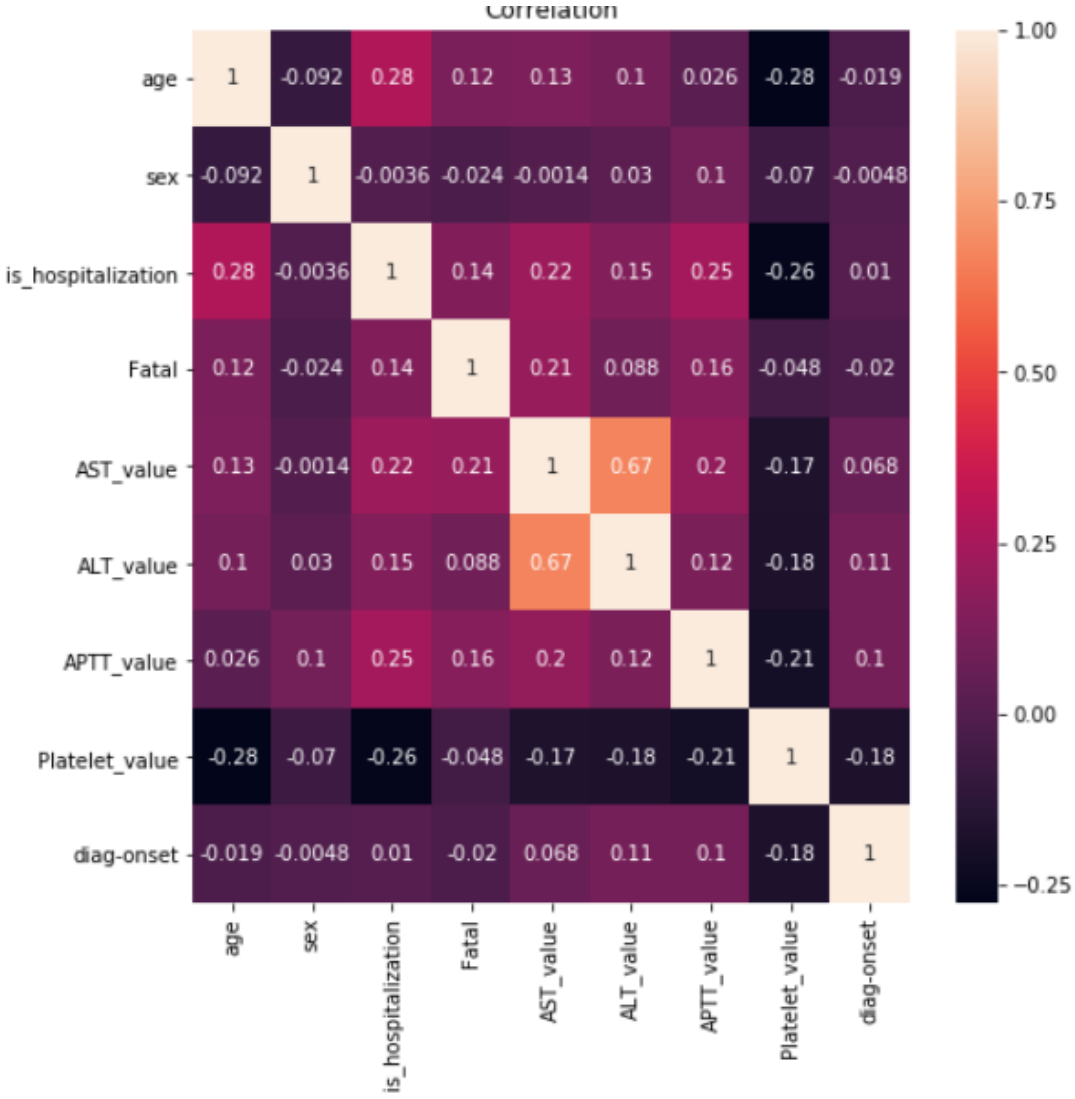
特徵欄位左至右依序為:年齡、性別、住院與否、AST 指數、ALT 指數、APTT 指數、Platelet 指數、確診減去發病天數。

指標欄位為:致命與否(Fatal)。

- 資料分析:

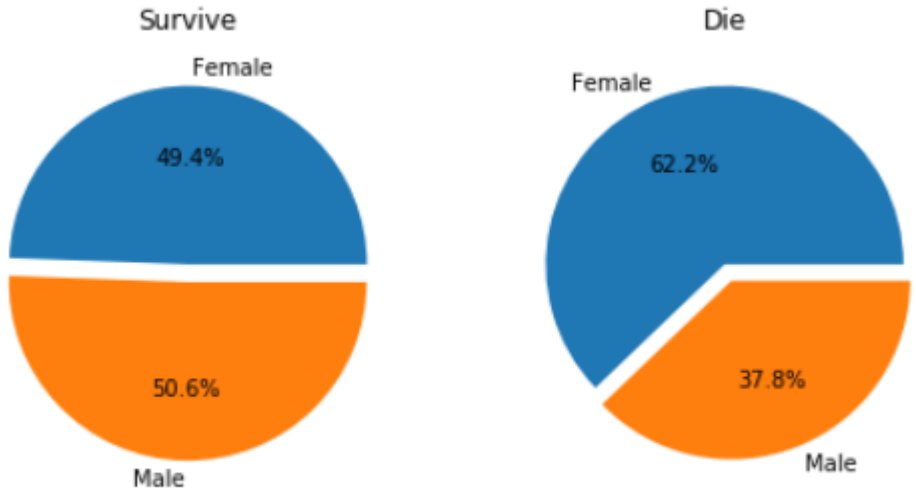
透過資料視覺化，判斷及分析各特徵與致命與否的關聯。

➢ 相關性分析，以熱圖呈現:



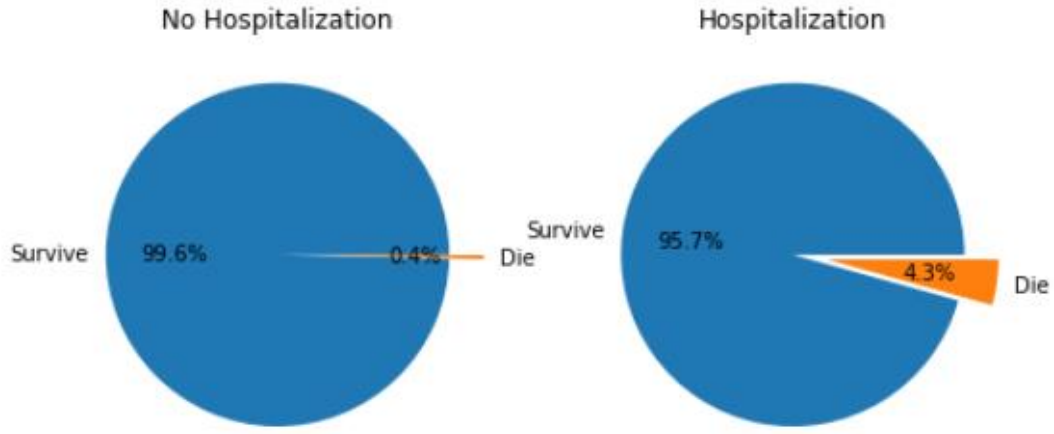
接著從相關度較高的特徵中，一一提取出來視覺化，觀察趨勢。

- 性別與致命與否的關聯圖:



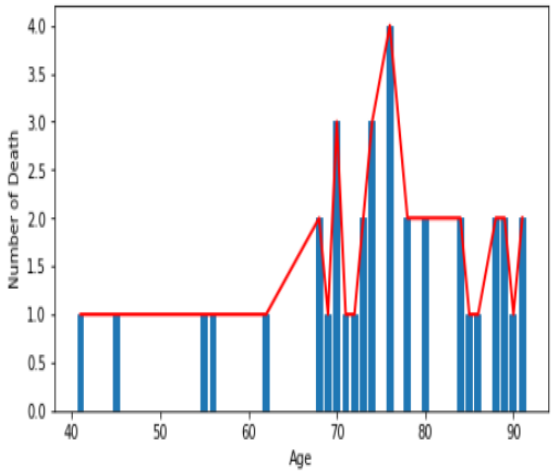
可看出，男女之間得病機率幾乎相同，但是女性死亡比例稍高。

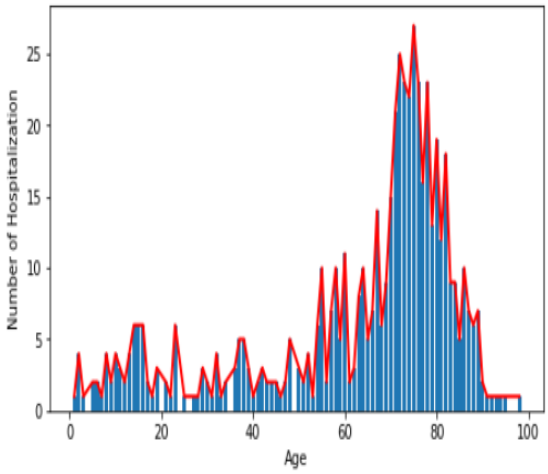
- 住院與否與致命與否的關聯圖:



未住院的患者中幾乎無人死亡，推測是因為醫師準確判斷病情並不嚴重，故未安排住院；住院的患者相對危險，死亡比例也就較高。

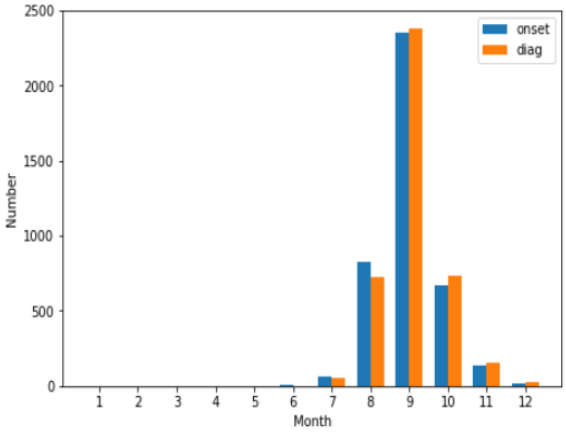
➢ 年齡與住院與否及死亡與否的關聯圖:

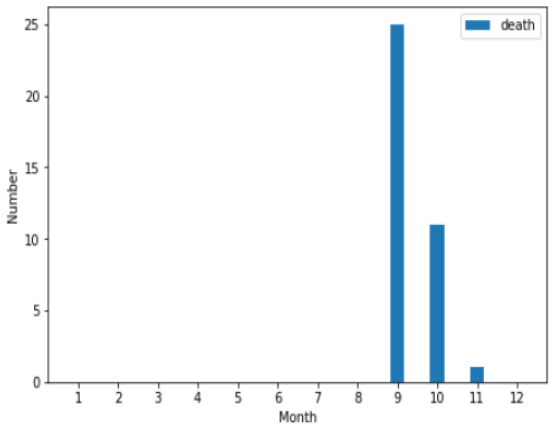




可以合理推測年齡越高，患病後的風險也就越高；故不論是死亡或住院的曲線都偏右，也就是年齡高的族群。

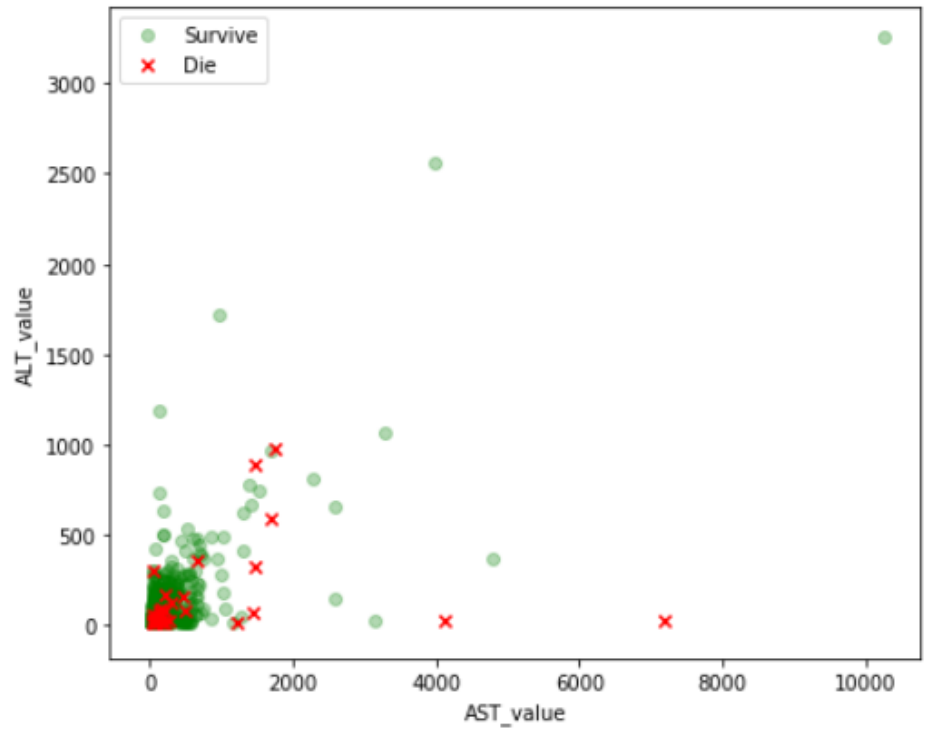
➢ 登革熱好發時期:

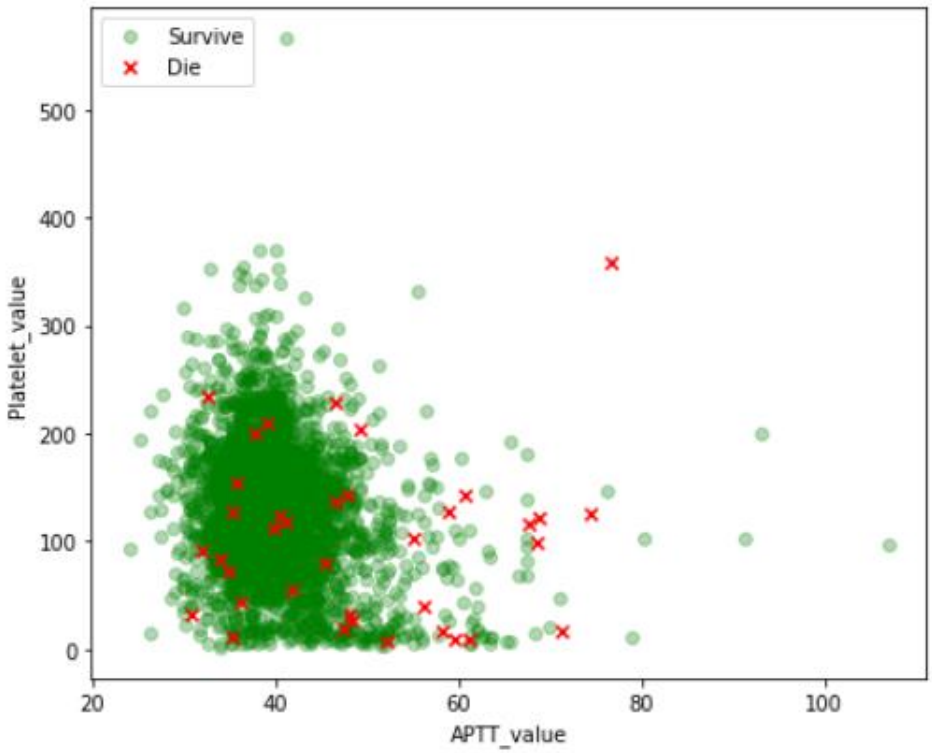




由圖可以看出登革熱流行的巔峰在 7 至 10 月之間，與一般人印象相符。

➢ 四個醫療指數與致命與否的關聯圖:



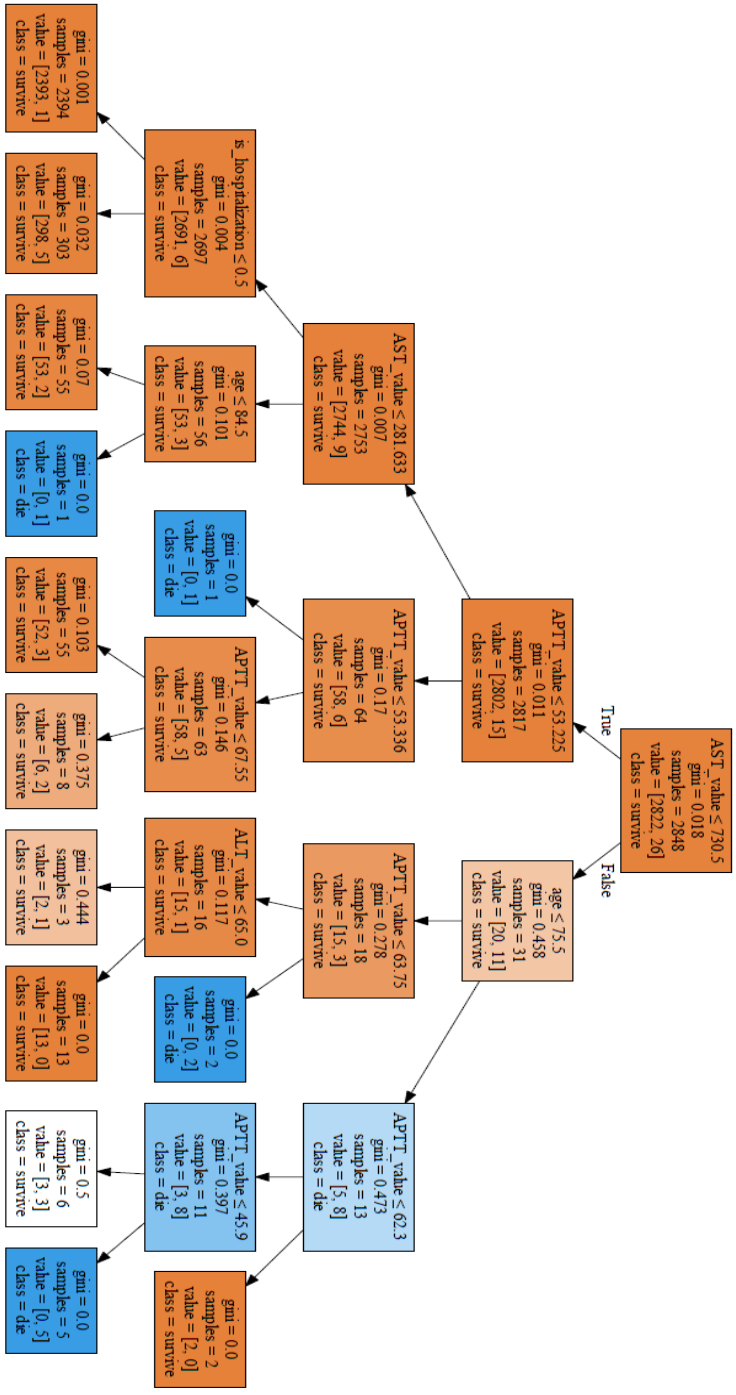


可以看出 AST 與死亡率相關性較高，下圖則顯示 APTT 的高相關性。

由上述資料分析步驟，可以看出特徵對於標籤值的影響程度及分布趨勢，接下來進行機器學習步驟:

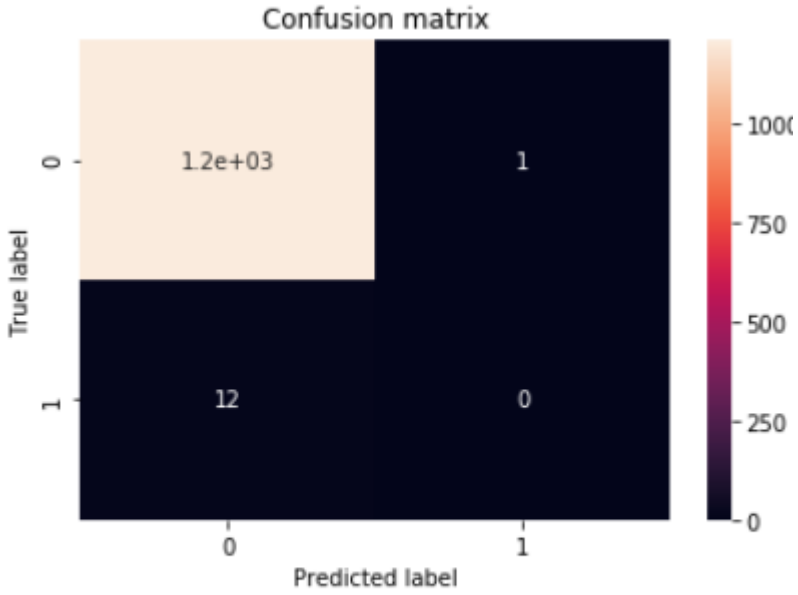
- **模型建立:**

  - 首先建立決策樹模型，可以看出決策判斷閾值。



模型預測準確度高達 99%，我們認為基於 baseline model，這樣準確度很高，但因為死亡樣本數少，故可能因此容易猜測，對此結果存疑，因此用混淆矩陣及 ROC Curve、AUC 面積來判斷模型價值。

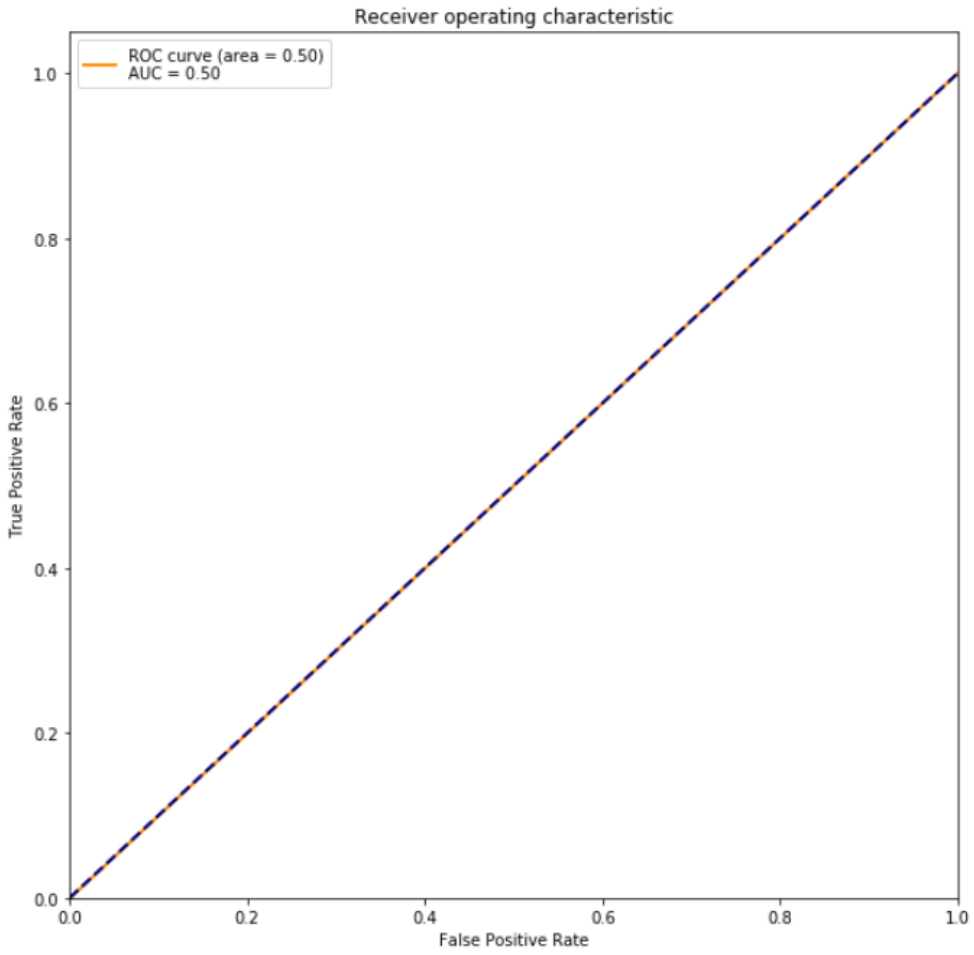
- 混淆矩陣:



	Precision	Recall	FPR	TPR	F1
0	99.917287	99.016393	100.0	99.016393	99.4648

可看出 FPR 竟高達 100%，初步判斷此模型訓練成效並不佳。

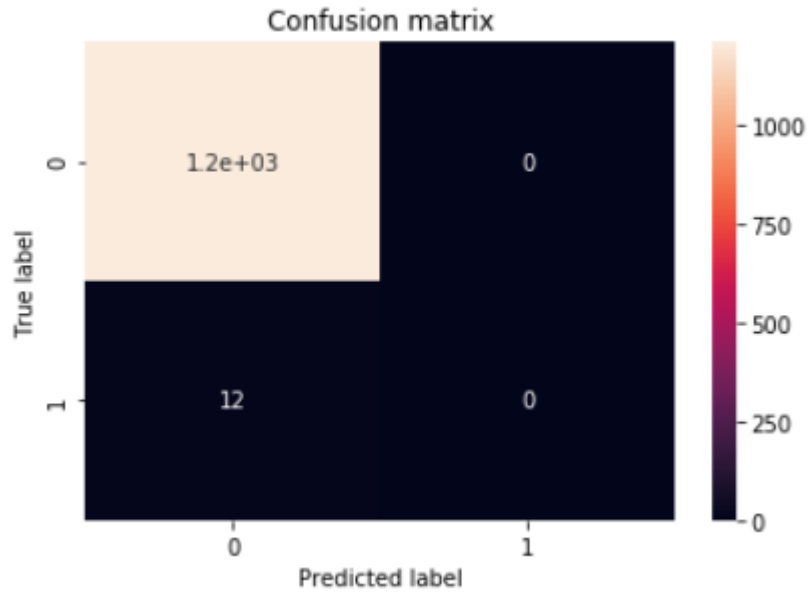
- ROC Curve&AUC:



輔以 ROC 曲線判斷，可知此決策樹模型的確成效不彰。

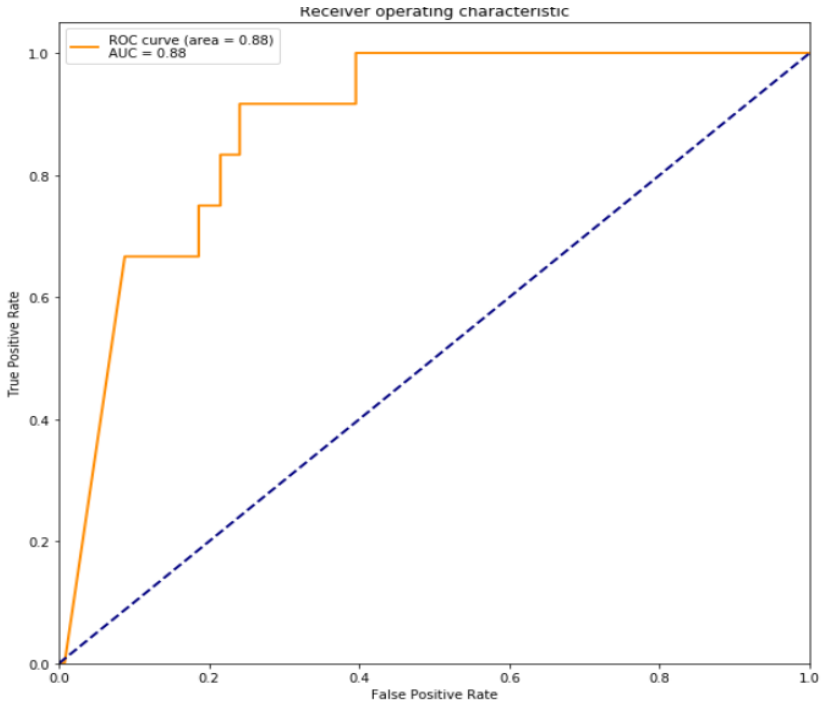
- 建立 SVM 二元分類模型:

- 混淆矩陣:



	Precision	Recall	FPR	TPR	F1
0	100.0	99.017199	NaN	99.017199	99.506173

可看出頗具成效。接著觀察 ROC 曲線及 AUC。



從中看出分類效果也不錯。應有潛力作為評斷標準。

結論

此次專案遭遇較大的問題，是空缺值過多及死亡筆數過少。第一個問題因原先採用平均數及中位數填值，但發覺會不符合真正的趨勢，如死亡病歷的醫療指數會特高或特低，但用平均數及中位數則無法展現。因此決定採取隨機森林回歸法，藉由觀察有完整資料的病歷資料樣本，來回歸出其餘空缺值，使較符合趨勢。

第二個問題則較難處理，死亡筆數過少，導致機器學習成效不彰，即使全部都猜存活，仍會有高達 98%的準確率。但若為了提高訓練成效，而在資料集增加死亡筆數，又恐有醫療專業上的疑慮，可能加入同個時間軸時的外部同屬性資料，提高樣本數，但因屬醫療機密，故仍十分困難。

因此此次專案，於機器學習預測的部分，應僅供參考，但仍從初期的資料處理及分析，整理出一些病理上的趨勢，如女性死亡率稍高、住院病患死亡率較高、年齡越高死亡率越高、AST 與 APTT 數值與死亡率的相關性。

未來若能應用決策樹或其他模型，透過學習預測，提出更多醫療指數的閾值(如 AST 大於 1000 則極度危險)，這種量化的數據能實質地院方判斷，則實用價值將大大提高。