登革熱資料分析（2015 台南地區登革熱疫情資料）

組名：FDD2

組員：黃彥瑋，蔡明修

簡介：

此資料集是 2015 年台南地區爆發大規模登革熱疫情，醫院所做的統計資料

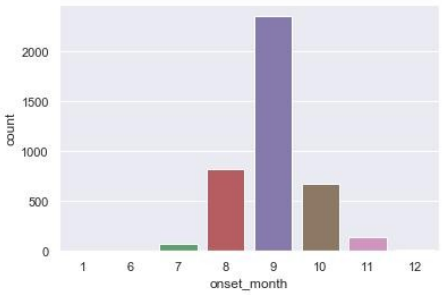
分析資料：

拿到的檔案共有四個 excel 檔，這是第一個，關於病人的一些基本資料

觀察圖片就醫日期分佈可以發現登革熱多集中在較熱的 8、9、10 月

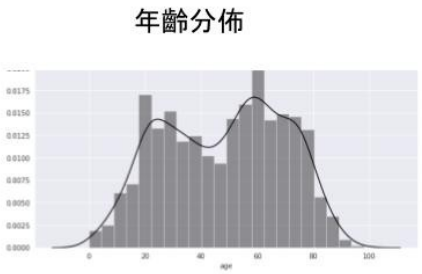
欄位說明

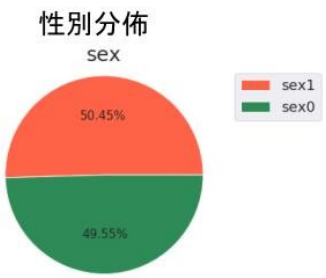
chartno：病人編號
age：年紀
sex：性別
onset_date：發病日期
diag_date：診斷日期
death_date：死亡日期（若病人沒有死亡則為NaN）
is_hospitalization：病人是否住院
Fatal：病人是否死亡（1：死亡、0：活著）

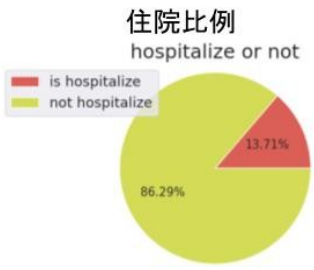


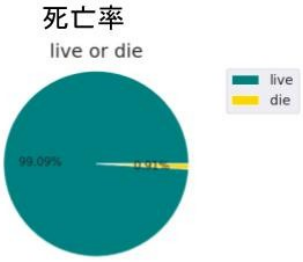
	chartno	age	sex	onset_date	diag_date	death_date	is_hospitalization	Fatal
0	A1564	74	1	2015-08-31	2015-09-02	NaN	0	0
1	A1878	71	1	2015-09-09	2015-09-15	NaN	0	0
2	A8146	38	0	2015-08-11	2015-08-14	NaN	0	0
3	A8476	55	0	2015-09-17	2015-09-17	NaN	0	0
4	A15171	44	1	2015-09-28	2015-09-28	NaN	0	0

進一步分析年齡、性別、住院比例和死亡率









另外四個資料集，分別是病人 AST、ALT、APTT、Platelet 的檢測數據。
AST 和 ALT 就是我們俗稱的肝指數，APTT 和血小板都和凝血有關

AST：天門冬氨酸轉氨酶

	chartno	type	Day	value
2816	A10015442	1	4	39
2817	A10017629	1	3	58
2818	A10017629	1	6	61
2819	A10030438	1	5	171
2820	A10030438	1	7	132

ALT：血清轉胺酶

	chartno	type	Day	value
0	A8130786	1	0	18
1	A8152157	1	1	14
2	A8152157	1	6	62
3	A17181845	1	5	77
4	A17181845	1	3	14

type：是否死亡（死亡為0）

Day：第n天量測的結果

value：濃度

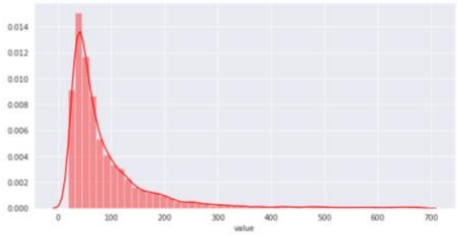
APTT（Activated partial thromboplastin time）：部份凝血活酵素時間

	chartno	type	Day	value
0	A8152157	1	1	37.9
1	A17189166	1	5	56.7
2	A12555309	1	4	41.4
3	A10852588	1	2	44.3
4	A10852588	1	7	37.8

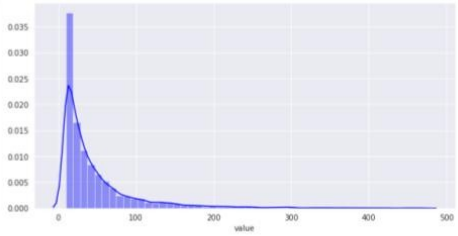
Platelet：血小板

	chartno	type	Day	value
0	A8130786	1	0	141
1	A8152157	1	6	79
2	A10323663	1	5	120
3	A10323663	1	1	199
4	A7214571	1	4	146

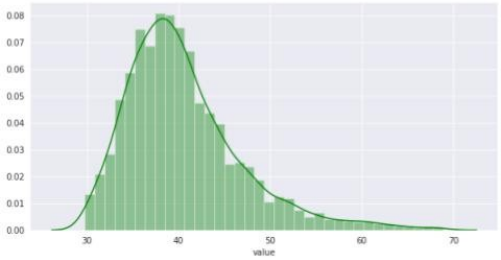
AST：天門冬氨酸轉氨酶



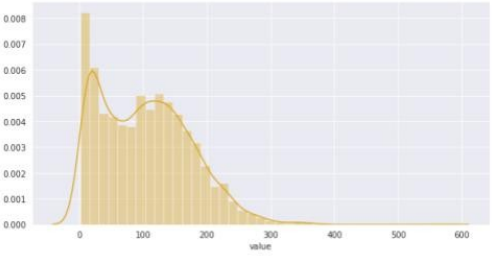
ALT：血清轉胺酶



APTT（Activated partial thromboplastin time）：部份凝血活酵素時間



Platelet：血小板



設定問題：

拿到一個疾病的資料，首先很直覺的，我們想預測病人死亡率，但觀察後發現死亡的人數極低（死亡率 < 1%），我們很難從死亡病患的資料做分析，且資料完整性不夠，不是每個病人都有做四個數值（AST、ALT、APTT、Platelet）檢測，因此預測死亡做出來的結果意義不大。

所以，我們另外訂了一個有實用性的題目：探討病患住院受什麼因素影響，並做預測。
我們針對病患基本資料和四種數值檢測，預測病人需不需要住院

把分析分成主要的兩個項目：基本欄位和物質濃度

分析分成兩個部分：

1. 基本欄位：

	chartno	age	sex	onset_date	diag_date	death_date	is_hospitalization	Fatal
0	A1564	74	1	2015-08-31	2015-09-02	NaN	0	0
1	A1878	71	1	2015-09-09	2015-09-15	NaN	0	0
2	A8146	38	0	2015-08-11	2015-08-14	NaN	0	0
3	A8476	55	0	2015-09-17	2015-09-17	NaN	0	0
4	A15171	44	1	2015-09-28	2015-09-28	NaN	0	0

2. 物質濃度：

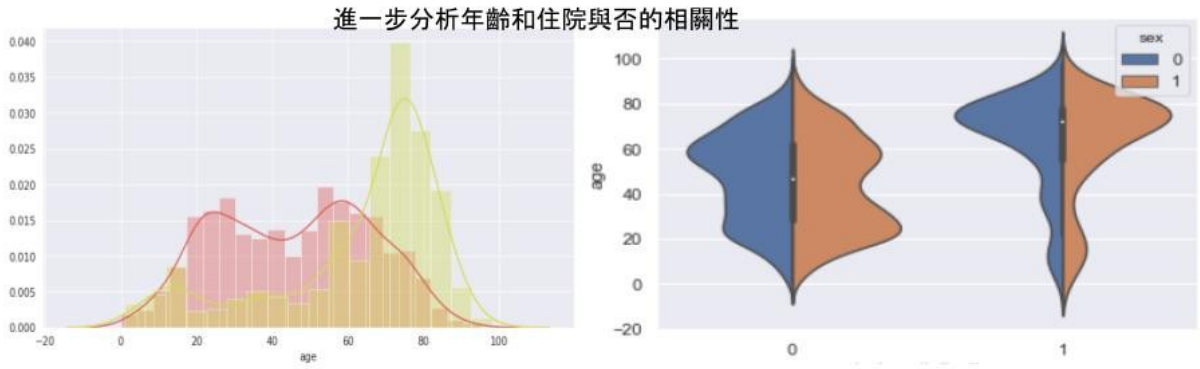
	chartno	type	Day	value
0	A8130786	1	0	141
1	A8152157	1	6	79
2	A10323663	1	5	120
3	A10323663	1	1	199
4	A7214571	1	4	146

基本欄位：

先用圖片觀察一下住院和那個 attribute 關係最深，發現是年齡

分析住院機率和哪個attribute關係最深，發現是年齡

是否住院和性別關係不大



is_hospitolization v.s. age
yellow: 住院
red: 未住院

is_hospitalization v.s. age, sex

- age大住院的比例較高
- 和sex 沒有明顯的關係

住院比例和年紀有關，用平均年紀(48歲)把dataset一分為二，發現
>48歲為20.82%，<48歲為5.92%

建立模型預測，發現各種模型數據差距不大，準確率都在 87%左右

預測是否住院

用最初的total dataset

Baseline：全部預測不住院，準確率為86.49%

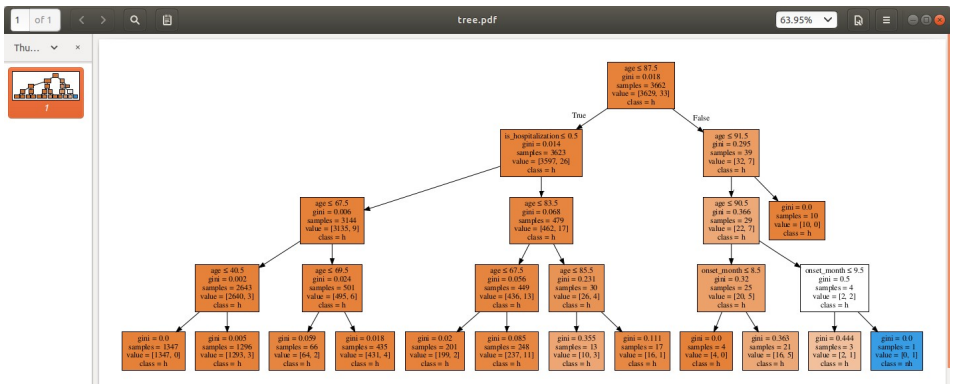
Decision Tree：經測試，最好的結果都是在87.73%，樹高度為5

SVM：86.48%

Logistic Regression：86.48%

features selection: age, (sex, onset_month, diag_month)

附上 decision tree 結構圖



物質濃度:

首先觀察血小板（Plateler）和住院的關係

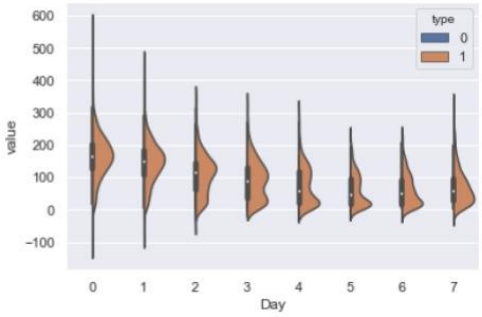
發現有住院的病人血小板濃度較沒住院的病人低

Platelet

隨著天數增加

Platelet濃度整體有下降的趨勢。

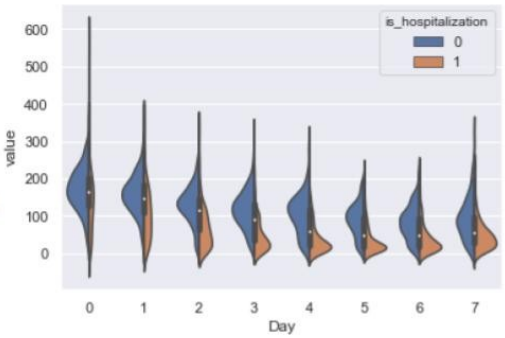
type = 0人數太少看不出來



Platelet

將violin plot左右改成is_hospitalization

- 隨著天數增加無論是否住院Platelet都有下降的趨勢
- 有住院的病人Platelet濃度整體較沒住院的低



so，用血小板預測是否住院!

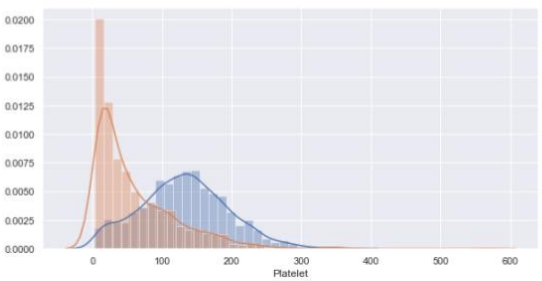
建立 Decision tree 並預測，準確率 83%，好像不是很優

Platelet

Platelet value越小住院機率越大

Platelet_df = pd.merge(Platelet, df)

Decision Tree: 83.01%

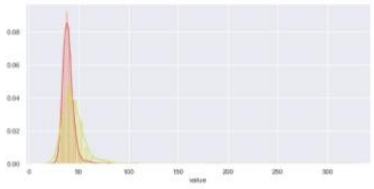


橘:有住院

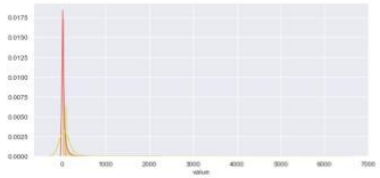
藍:沒住院

將其他三種濃度也拿來預測，結果如圖
和血小板相反，AST、ALT 和 APTT 都是濃度越高，住院機會越高
但用決策樹做預測後，準確率也都不是太高，座落在 77~82%左右

APTT

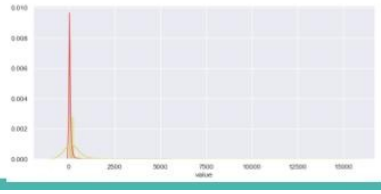


ALT

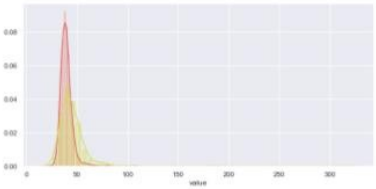


和Platelet相反
value大住院機率較高
且沒有Platelet顯著

AST

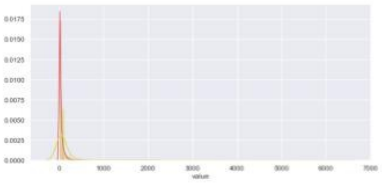


APTT



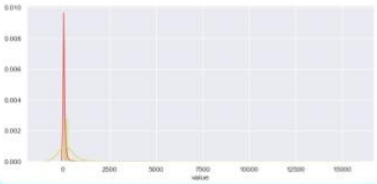
Decision Tree：82.45%

ALT



Decision Tree：77.78%

AST



Decision Tree：81.71%

因為準確率一直沒達到預期，我們思考要怎麼預測的更好，想到因為我們不是本科系，對這些濃度數據不熟，不知道他們的值對應了什麼訊息。
因此我們查了關於這些數據的資料，發現了這些濃度資料和登革熱的關係，且發現醫院會把 **AST/ALT** 當作一項指標!

Survey data

SEARO Dengue Guideline 2011《登革熱/登革出血熱 臨床症狀、診斷與治療》疾病管制署 2013

登革出血熱/休克症侯群 – 實驗室檢查

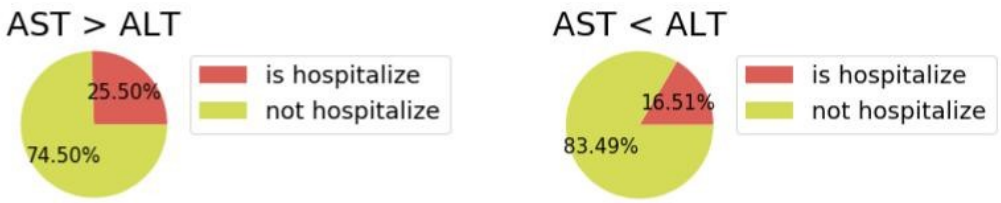
- 白血球低下 (< 5000 cells/mm^3)
- 血小板低下 (< 10萬 cells/mm^3)
- 肝功能上升
  - AST ≤ 200 U/L
  - AST/ALT > 2
- PT, aPTT延長
- 低血鈉、低血鈣
- 代謝性酸血症
- ESR不高(可用以區別一般敗血性休克)

https://www.health.taichung.gov.tw/media/365041/%E7%99%BB%E9%9D%A9%E7%86%B1%E9%98%B2%E6%B2%BB%E6%95%99%E8%82%B2%E8%A8%93%E7%B7%B4%E8%AC%9B%E7%BE%A9.pdf

---

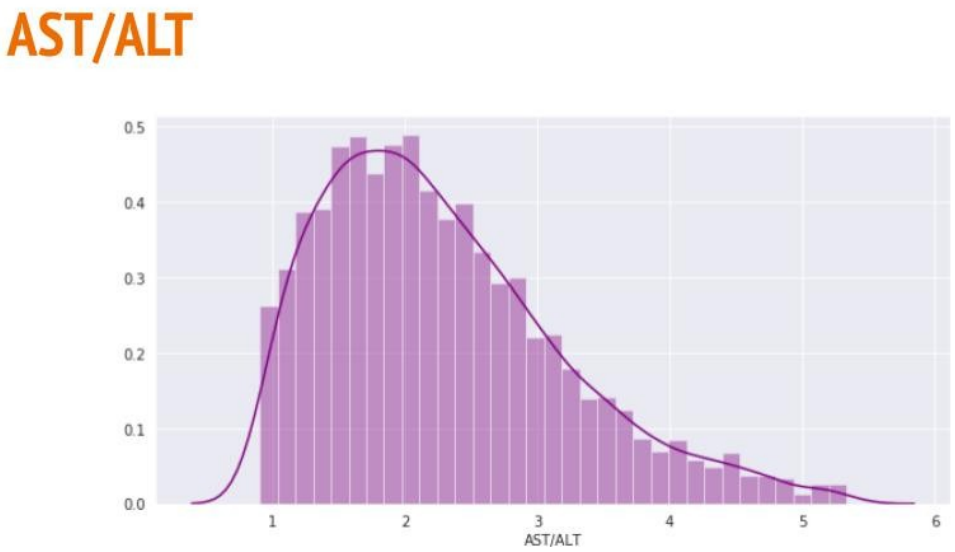
得到這些寶貴的資訊後，我們首先觀察 AST、ALT 之間相對大小和是否住院的關係
發現 AST>ALT 時住院機率大於 AST<ALT 的住院機率!

比較AST, ALT相對高低 和 住院與否的關係



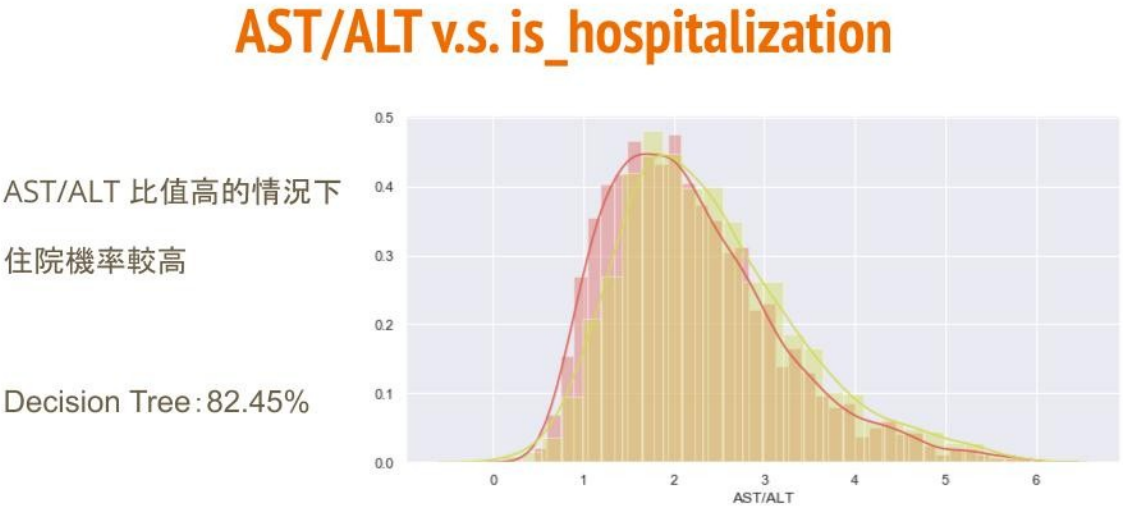
AST > ALT 住院機率較高

前面得知 AST/ALT 有參考性，因此觀察一下它的分佈



觀察其和是否住院之關係，發現確實當 AST/ALT 較高時住院機率高!
很高興的拿他來預測，以為這樣結果肯定不錯

BUT
用決策樹做出來，準確率只有 82.45% /_\
不過稍微觀察圖片也會發現 AST/ALT 雖然會影響住院機率，但關聯性似乎也不是那麼大



結論＆心得：

相對於**年齡**，體內物質的濃度對於住院的影響應該不如我們所想的那麼大，
且因為此**資料集並不完整**（四種濃度檢測不是每個病人都有做，有些病人只檢測一兩種，有些完全沒有檢測），導致 merge dataset 時數據量會顯著變少。
此資料集不是很大，只有幾千筆，且並不是很完整，我們只能從現有的資訊盡量做分析，但放到現實中，我們會遇到的資料集或許就是長成這個樣子，所以我們覺得這次的 project 是一個學習資料分析上很好的經驗，學習從不完整的資料得到有意義的結論。

為什麼accuracy 反而降低了？

1. 住院與否和體內物質濃度沒那麼相關
2. merge dataset 的問題
3. age 影響力極大

Features importancy

有age[0.84147222 0.01882717 0.01924477 0.03067823 0.02572237 0.06405524]

age AST/ALT

AST/ALT

沒age[0.04616311 0.06862451 0.13670197 0.22448374 0.52402667]

age 的影響比起Platelet, AST.....等還是太高了

Age matter!

Conclusion

Age matters!