資料探勘 HW2 報告

報告說明：

這次的作業主要藉由自己產生 Data 與設定 Rule 替 Data 做分類標示,再觀察不同的分類器是否能正確地找出 Rule 來分類 Data,並觀察分析若 Data 有雜訊的結果下是否有不同結果。

報告架構：

1. 題目設定 & Data design
2. Features
3. Absolute right rules
4. Data analysis with decision tree model
5. Data analysis with SVM model
6. Data analysis with KNN model

題目設定 & Data design

題目設定：

評估 40 歲以上有急性肩膀疾病的患者是否也有併發早期五十肩(Adhesive capsulitis)。

Motivation：

肩膀疾病在現今社會中非常常見，可能原因有夾擠、滑囊炎、肌腱炎、肩關節囊唇撕裂...等等。若在初期沒有善加處理，忍耐著疼痛不進行復健可能會演變為五十肩。但五十肩有時會被誤診為其他肩膀疾病，其中最常被誤診為肩膀滑囊炎。因此希望可以藉由一些理學檢查作為特徵，決定此患者是否患有早期五十肩。

Data Design：

由於五十肩好發在 40 歲以上的族群，因此本次資料年齡會限定在 40 歲以上的 的病患，並模擬實務上遇到肩膀疼痛病人會做的理學檢查與病患自述症狀來評估病患是否有早期五十肩的狀況。

在 Data 的設計上面，為了更貼近現實狀況，對於 Data 做出以下限制。

例如：

A. 若有糖尿病，病患的 ROM 通常會受到影響，平均是較一般人下降 5°。(Tiffany K. Gill,2022)

B. AROM<= PROM (AROM 可能受到肌力、肩膀是否有受傷或是附近肌肉是否過度緊繃等等因素影響，設計 AROM 有 20%的機率可能會 < PROM，80%的機率 = PROM，通常一般人 AROM = PROM)

C. 若 Active Flexion ROM < 90°，則會無法實施 Empty Can Test。此時 Empty Can Test 會標示為 -1 表示無法測試。

D. 臨床上五十肩患者發生率約 3~5%，為了使產生的 dataset 接近臨床數據，因此也會將五十肩病患產生率控制在 3~5%

Noise：

由於五十肩偶爾會被誤診為滑囊炎(Bursitis) 或是沒有做 Coracoid process test 將假性五十肩(muscle guarding 造成)誤診為五十肩，因此做 noise 資料分析時設定會有 0.08 的機率將假性五十肩診斷為五十肩，0.05 的機率將滑囊炎診斷為五十肩，而五十肩則有 0.05 的機率被診斷為沒有五十肩，模擬含有臨床上誤診的情況收集的 dataset。

Features

Attributes	Possible values	Description
	Base Information	
Age	40~100	Age of the patient Generated randomly
Gender	Female(1) or male(0)	Gender of the patient Generated randomly
Weight	Positive number	Weight of the patient Generated with Normal distribution
Height	Positive number	Height of the patient Generated with Normal distribution
BMI	Positive number	BMI of the patient
	Passive Shoulder ROM & Active Shoulder ROM test	
Active Shoulder Flexion ROM	0~180°	Active Flexion ROM of the patient Generated with Normal distribution
Active Shoulder Extension ROM	0~50°	Active Extension ROM of the patient Generated with Normal distribution

Active Shoulder Abduction ROM	0~180°	Active Abduction ROM of the patient Generated with Normal distribution
Active Shoulder External Rotation ROM	0~90°	Active External Rotation ROM of the patient Generated with Normal distribution
Active Shoulder Internal Rotation ROM	0~70°	Active Internal Rotation ROM of the patient Generated with Normal distribution
Passive Shoulder Flexion ROM	0~180°	Passive Flexion ROM of the patient Generated with Normal distribution
Passive Shoulder Extension ROM	0~50°	Passive Extension ROM of the patient Generated with Normal distribution
Passive Shoulder Abduction ROM	0~180°	Passive Abduction ROM of the patient Generated with Normal distribution
Passive Shoulder External Rotation ROM	0~90°	Passive External Rotation ROM of the patient Generated with Normal distribution
Passive Shoulder Internal Rotation ROM	0~70°	Passive Internal Rotation ROM of the patient Generated with Normal distribution
Tests & Subject Symptoms		
Normal on X-rays	Normal (1) or Abnormal (0)	X-rays of the patient Generated randomly
Coracoid Pain Test	Positive (1) or negative (0)	Coracoid Pain Test on the patient Generated randomly
Empty can test	Positive (1) or negative (0) or 無法測試 (-1)	Empty can test on the patient Generated randomly
Painful arc test	Positive (1) or negative (0)	Painful arc test on the patient Generated randomly
Have trauma on shoulder?	Yes (1) or No (0)	Does patient have trauma on shoulder? Generated randomly
Have difficulty sleep in night?	Yes (1) or No (0)	Does patient have difficulty sleep in night? Generated randomly
Upper arm pain?	Yes (1) or No (0)	Does patient have upper arm pain? Generated randomly
Pain when shoulder small movement?	Yes (1) or No (0)	Does patient have pain when shoulder small movement? Generated randomly
Underlying Disease		
Have diabetes?	Yes (1) or No (0)	Does patient have diabetes?

<p.s.> ROM (Range Of Motion)

Absolute right rules

1. 被動外轉(Passive External Rotation)角度 < 40°
2. 被動外展(Passive Abduction)或屈曲(Passive Flexion)角度< 125°
3. Coracoid Pain Test is positive
4. Pain in your shoulder even though smallest movement
5. Normal on X-rays

同時符合以上五個條件才會被分類為是五十肩病患

Data analysis with decision tree model

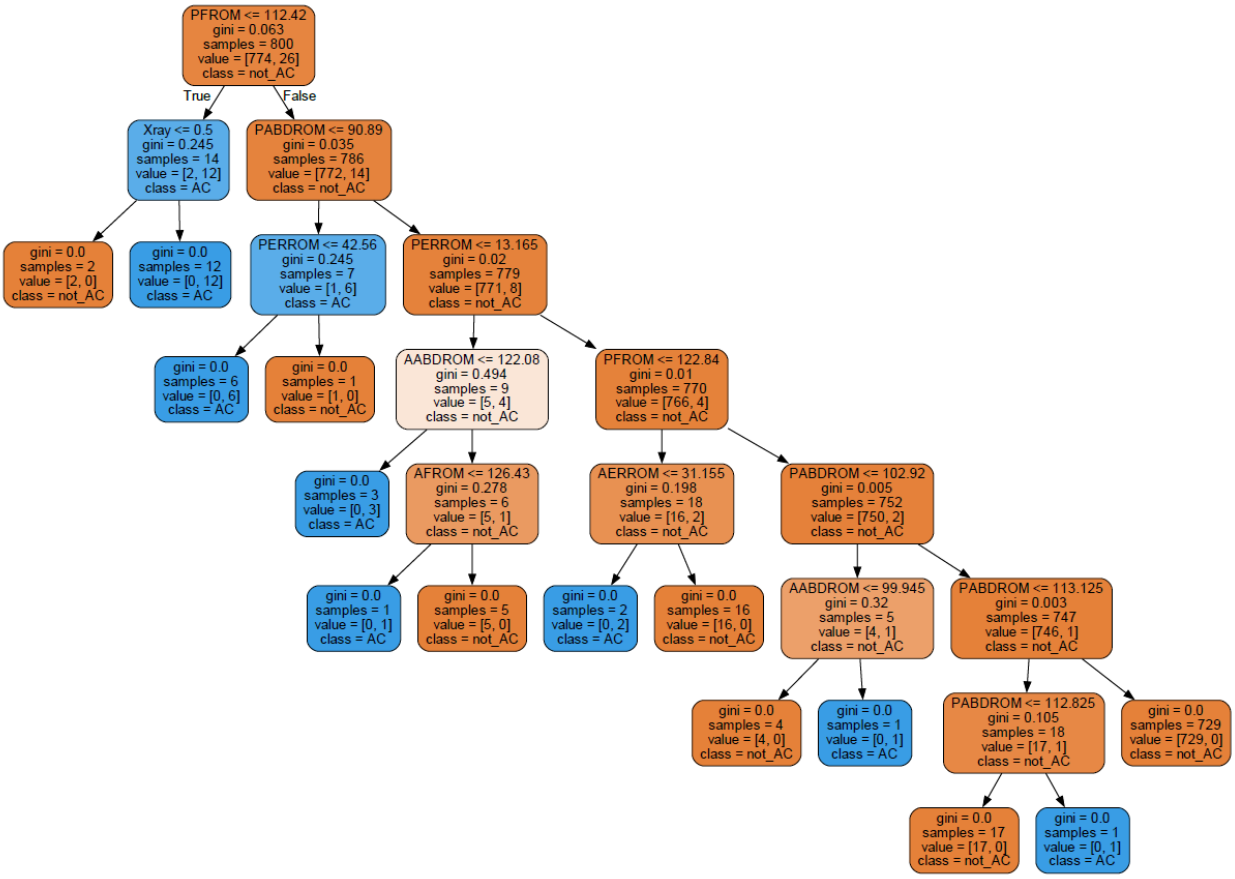
Classifier：sklearn.tree.DecisionTreeClassifier
說明：以下說明將會根據下列四種情況進行分析，並調整 Sample 的數量觀察是否得到不同的結果。

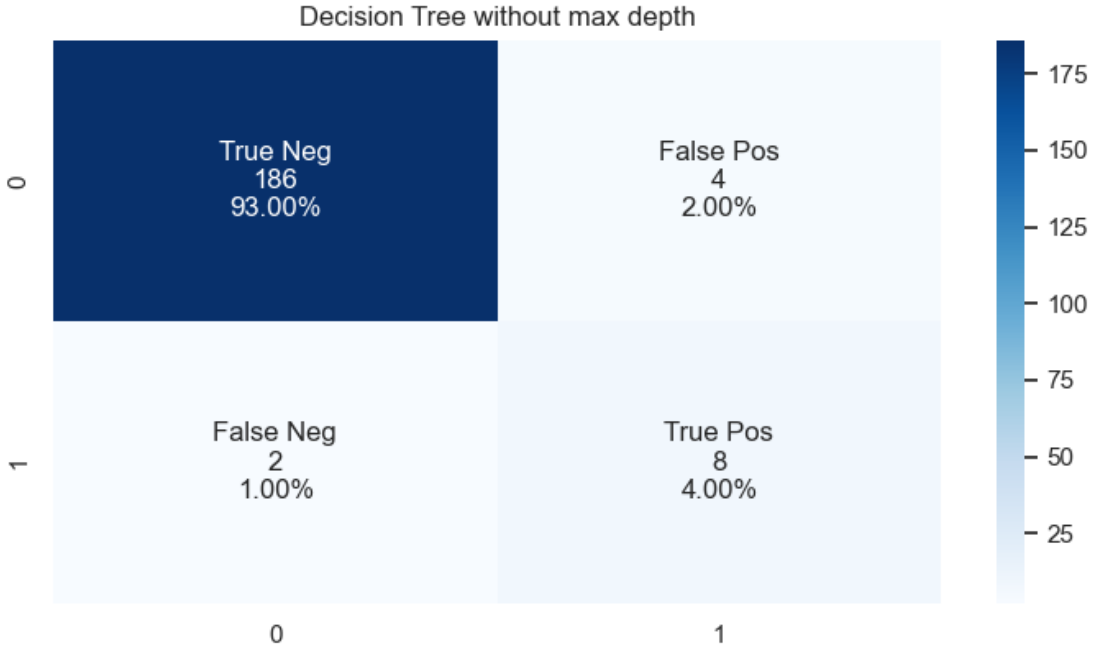
A. 不限制樹高，data 無 noise 也沒被標記錯誤(無誤診)
B. 限制樹高(max depth = 6)，data 無 noise 也沒被標記錯誤(無誤診)
C. 不限制樹高，data 有 noise 及有標記錯誤(有誤診)
D. 限制樹高(max depth = 6)，data 有 noise 及有標記錯誤(有誤診)

完全無誤診(無 noise)的情況

Dataset 1
Sample 數量：1000 (Training Data : Testing Data = 8:2)

A. 不限制樹高，data 無 noise 也沒被標記錯誤(無誤診)
   Tree height：7
   Accuracy：0.97 (6 筆資料分類錯誤)
   Precision：0.67
   Recall：0.8
   AUC：0.87


PFROM <= 112.42
gini = 0.063
samples = 800
value = [774, 26]
class = not_AC
True
False
Xray <= 0.5
gini = 0.245
samples = 14
value = [2, 12]
class = AC
PABDROM <= 90.89
gini = 0.035
samples = 786
value = [772, 14]
class = not_AC
gini = 0.0
samples = 2
value = [2, 0]
class = not_AC
gini = 0.0
samples = 12
value = [0, 12]
class = AC
PERROM <= 42.56
gini = 0.245
samples = 7
value = [1, 6]
class = AC
PERROM <= 13.165
gini = 0.02
samples = 779
value = [771, 8]
class = not_AC
gini = 0.0
samples = 6
value = [0, 6]
class = AC
gini = 0.0
samples = 1
value = [1, 0]
class = not_AC
AABDROM <= 122.08
gini = 0.494
samples = 9
value = [5, 4]
class = not_AC
PFROM <= 122.84
gini = 0.01
samples = 770
value = [766, 4]
class = not_AC
gini = 0.0
samples = 3
value = [0, 3]
class = AC
AFROM <= 126.43
gini = 0.278
samples = 6
value = [5, 1]
class = not_AC
AERROM <= 31.155
gini = 0.198
samples = 18
value = [16, 2]
class = not_AC
PABDROM <= 102.92
gini = 0.005
samples = 752
value = [750, 2]
class = not_AC
gini = 0.0
samples = 1
value = [0, 1]
class = AC
gini = 0.0
samples = 5
value = [5, 0]
class = not_AC
gini = 0.0
samples = 2
value = [0, 2]
class = AC
gini = 0.0
samples = 16
value = [16, 0]
class = not_AC
AABDROM <= 99.945
gini = 0.32
samples = 5
value = [4, 1]
class = not_AC
PABDROM <= 113.125
gini = 0.003
samples = 747
value = [746, 1]
class = not_AC
gini = 0.0
samples = 4
value = [4, 0]
class = not_AC
gini = 0.0
samples = 1
value = [0, 1]
class = AC
PABDROM <= 112.825
gini = 0.105
samples = 18
value = [17, 1]
class = not_AC
gini = 0.0
samples = 729
value = [729, 0]
class = not_AC
gini = 0.0
samples = 17
value = [17, 0]
class = not_AC
gini = 0.0
samples = 1
value = [0, 1]
class = AC


Decision Tree without max depth
True Neg
186
93.00%
False Pos
4
2.00%
False Neg
2
1.00%
True Pos
8
4.00%
0
1
0
1
175
150
125
100
75
50
25

B. 限制樹高(max depth = 6)，data 無 noise 也沒被標記錯誤(無誤診)

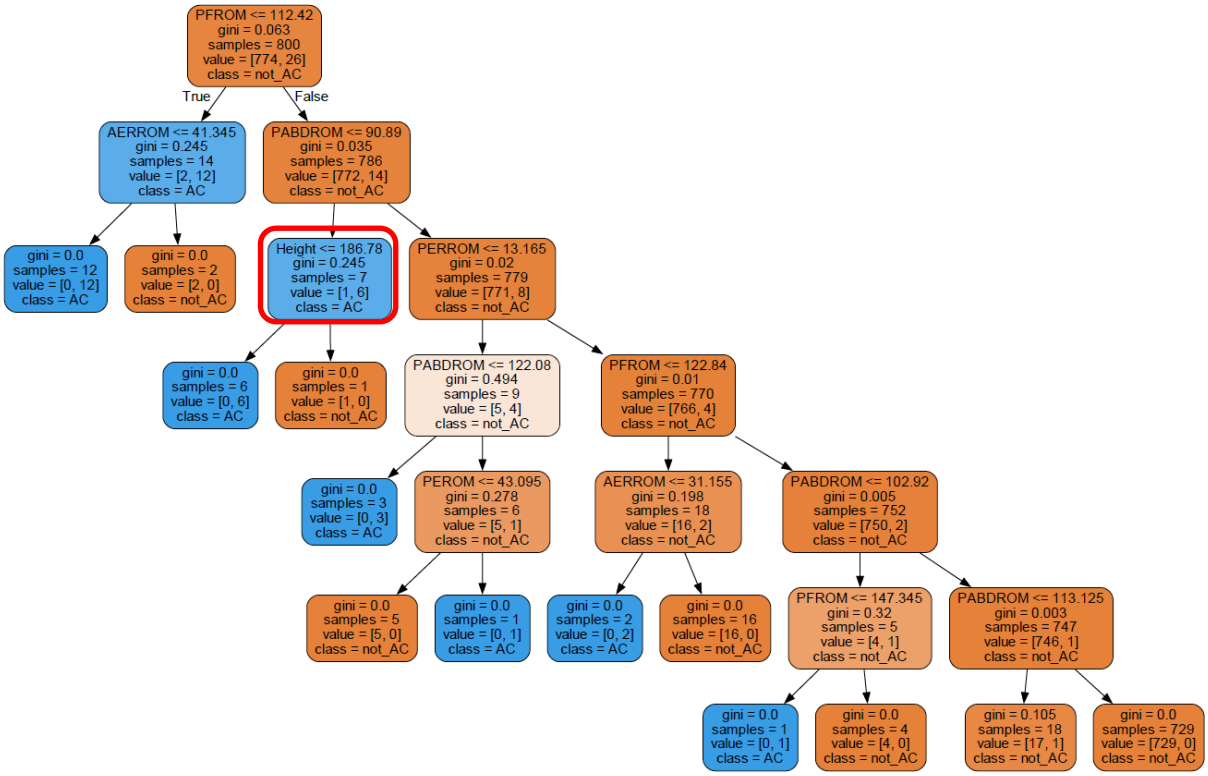
Tree height：6

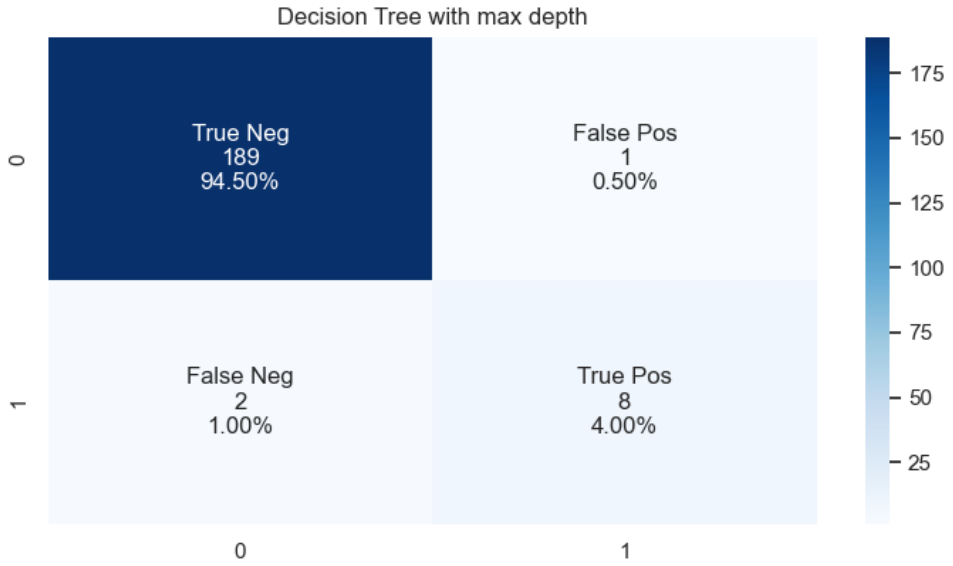
Accuracy：0.985 (3 筆資料分類錯誤)

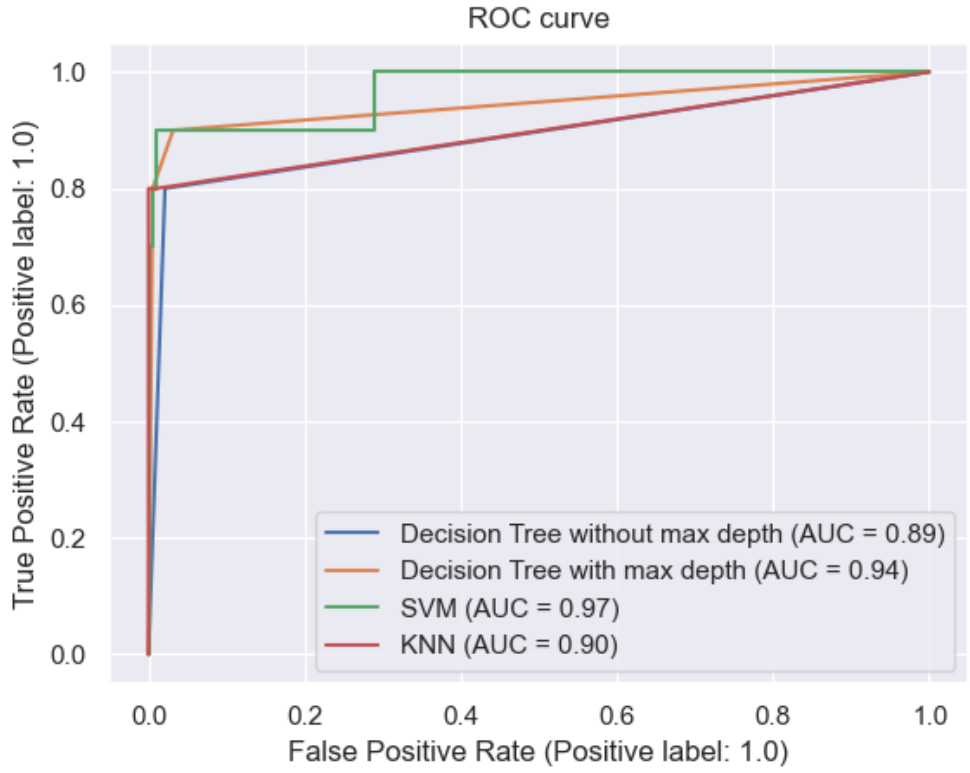
Precision：0.89

Recall：0.8

AUC：0.94







分析：

在 Dataset 1 中，雖然生成資料有 1000 筆，但其中 training data 被標記為有五十肩的病患(negative data)只有 26 位，或許是因為 positive data 數量和整體數量相比過少，才導致決策樹無法完全正確地找出規則進行分類。

在沒有限制樹高的決策樹模型(A 模型)中，有找出幾個接近 Absolutely right rule 的規則，或許是 data 數量只有 1000 筆相對較少，大部分找出的是錯誤的規則，雖然 Accuracy 到達 97%，不過 AUC = 0.87，Precision = 0.67，Recall = 0.8，似乎還有可以成長的空間。

在有限制決策樹高度的模型(B 模型)中，也是有找出幾個較為接近 Absolutely right rule 的規則，雖然不是完全正確，但相較於 A 模型，找出的規則大部分更接近 Absolutely right rule，因此在 dataset 1 的 Accuracy 是 B 模型較為準確。但 B 模型中還是有找出有點奇怪的分類規則，**例如：**身高居然成為分類的標準，或許也是因為 positive data 數量不夠多的關係才會有這種現象。和 A 模型相同，AUC、Precision、Recall 還有些可以成長的空間。

接下來我們將增加 Sample 數量進行測試。

Dataset 2

Sample 數量：10000 (Training Data : Testing Data = 8:2)

A. 不限制樹高，data 無 noise 也沒被標記錯誤(無誤診)

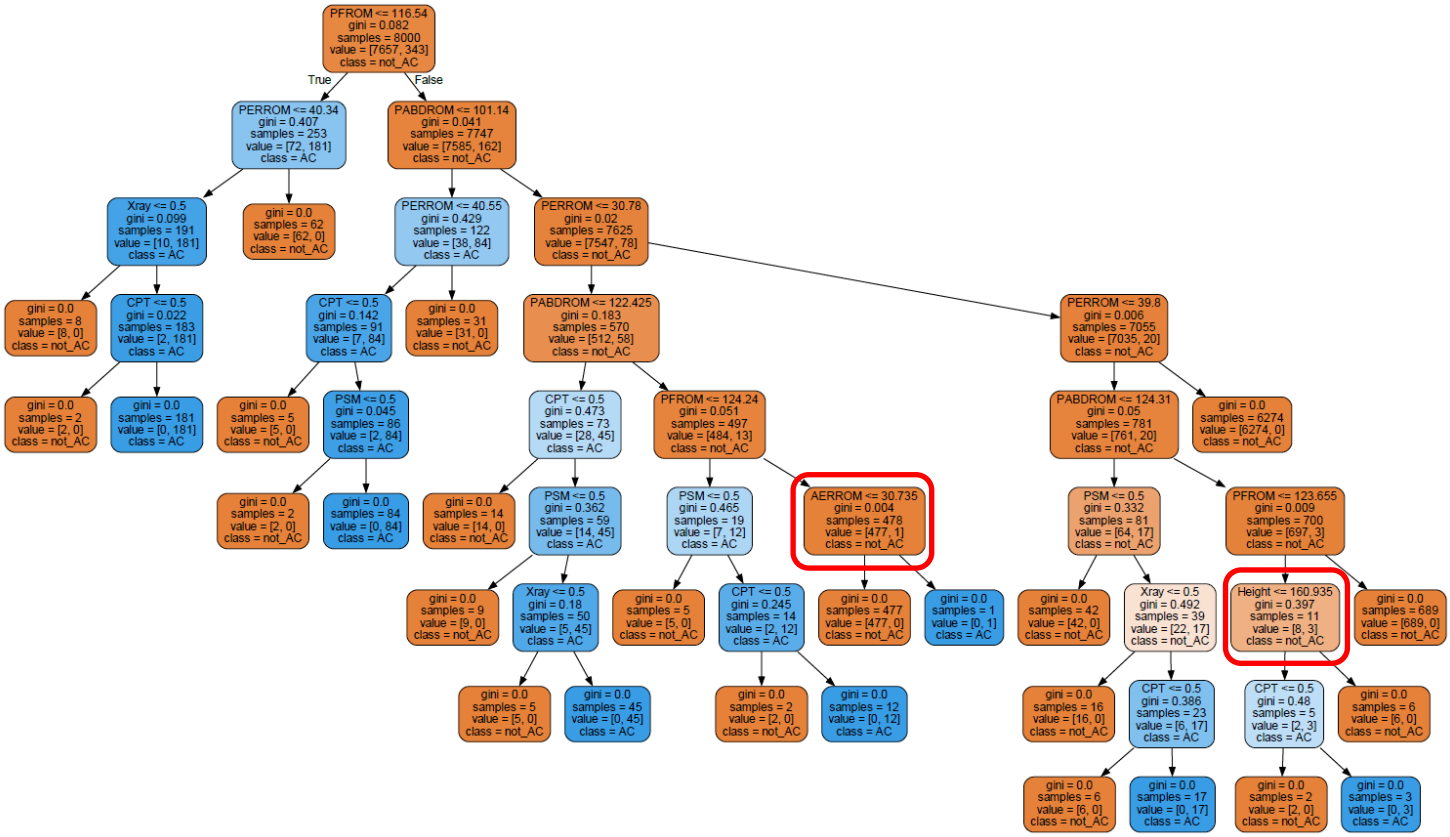
Tree height：8

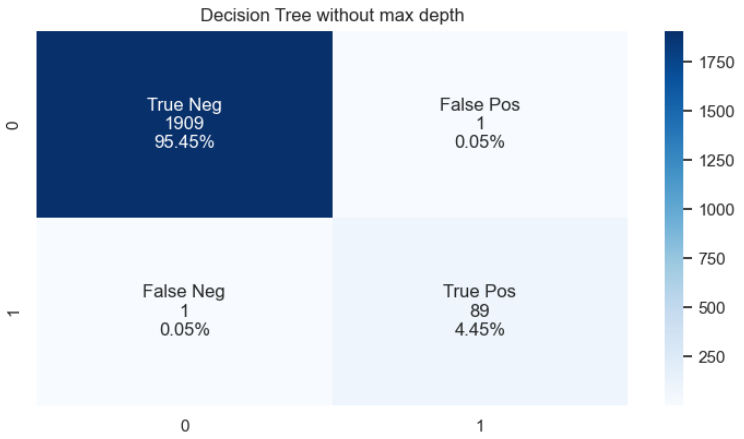
Accuracy：0.999 (2 筆資料分類錯誤)

Precision：0.989

Recall：0.989

AUC：0.99





B. 限制樹高(max depth = 6)，data 無 noise 也沒被標記錯誤(無誤診)

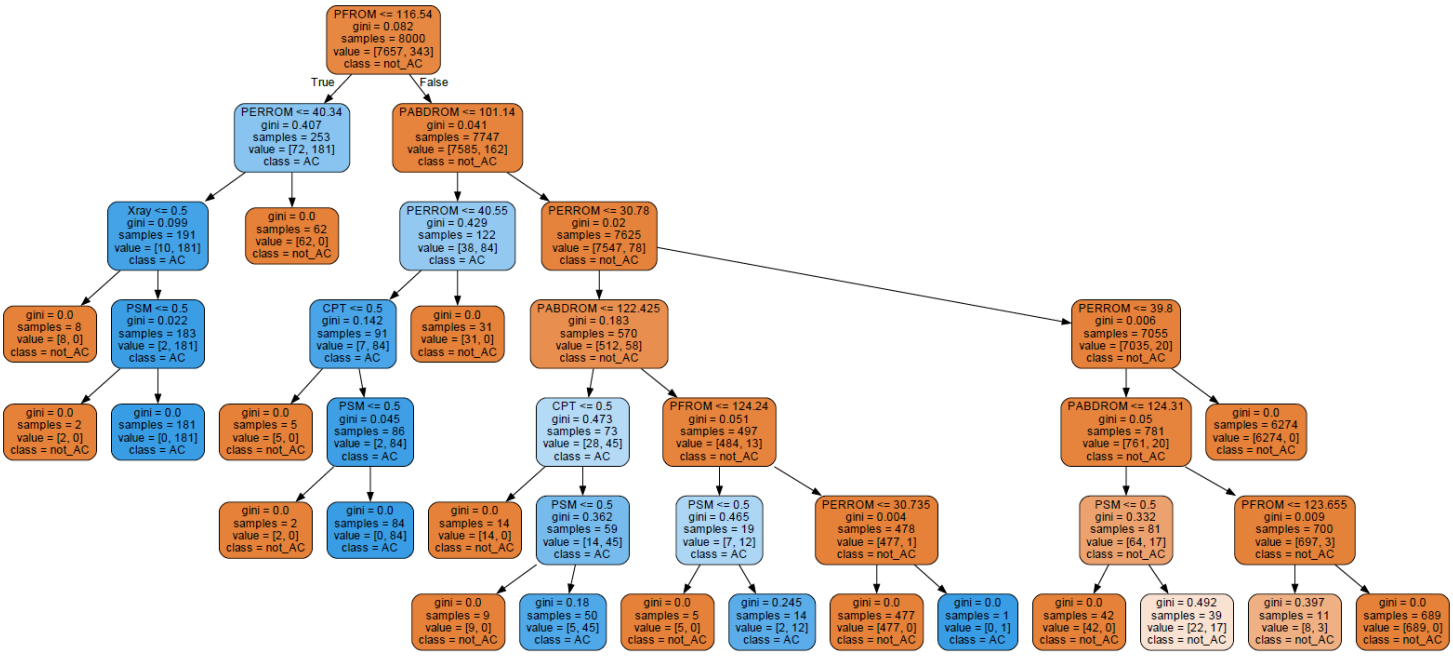
Tree height：6

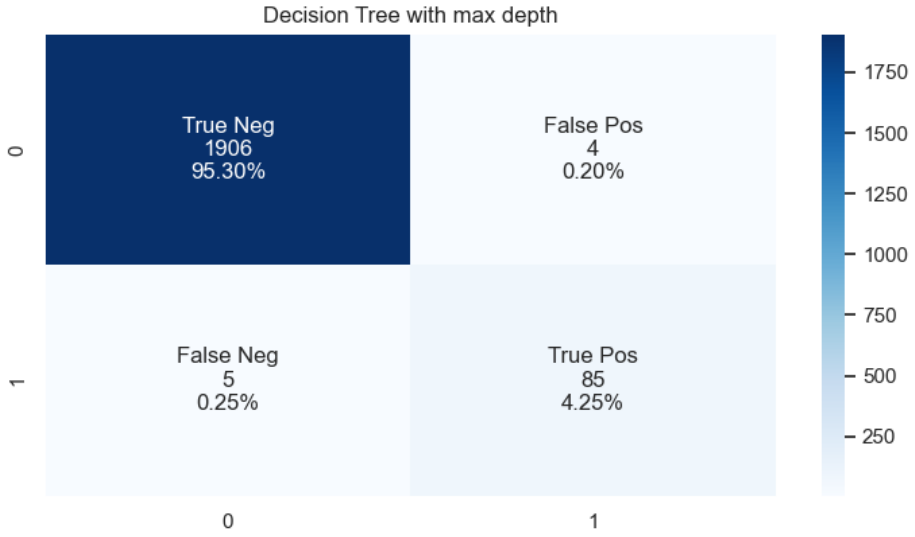
Accuracy：0.996 (9 筆資料分類錯誤)

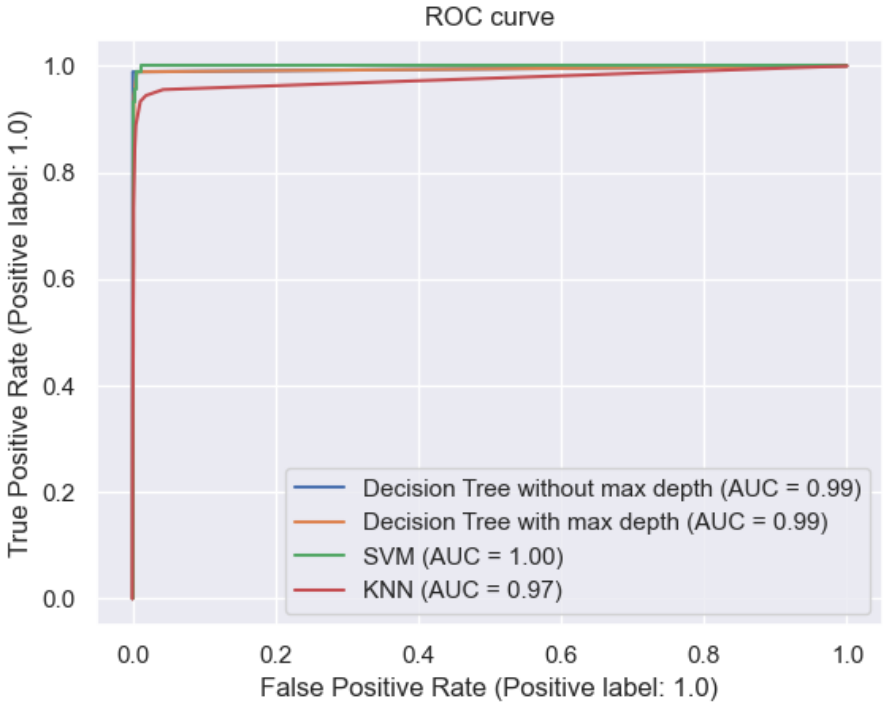
Precision：0.955

Recall：0.944

AUC：0.99







分析：

在 Dataset 2 中，生成資料有 10000 筆，這次不論有沒有限制樹高的決策樹，找到的幾乎都很接近 Absolutely right rule 的規則進行分類，甚至有幾條正是 Absolutely right rule。在 A、B 兩個模型中每條路徑都有找出 PERROM < 40 這個特徵作為 rule 判斷是否患有五十肩。在實際上的理學檢查中，這個對於評估病人是否為五十肩其中一項重要的依據，表示決策數有發現這個特徵是判斷的關鍵。雖然模型 A 整體 Accuracy、Recall、Precision 看起來較高，不過在模型 A 中還是有發現兩個較為奇特的規則，反倒是 B 模型所找出的規則都是在 Absolutely right rule 中。

例如：

A. A 模型中，規則中居然出現身高小於某數值會成為是否有五十肩判斷標準，這是相對較為奇怪的分類規則。

B. 另外一個比較特別的是其中有項規則為 AERROM>30 時會被判斷為有五十肩，出現這個結果似乎並不是這麼正確，或許剛好那條判斷路徑只剩下一個 positive data 要被分類，AERROM 可以剛好分類完畢，才會使用這個特徵作為分類標準。

因此接下來我們會再增加 Sample 數量進行測試。

Dataset 3

Sample 數量：100000 (Training Data : Testing Data = 8:2)

A. data 無 noise 也沒被標記錯誤(無誤診)

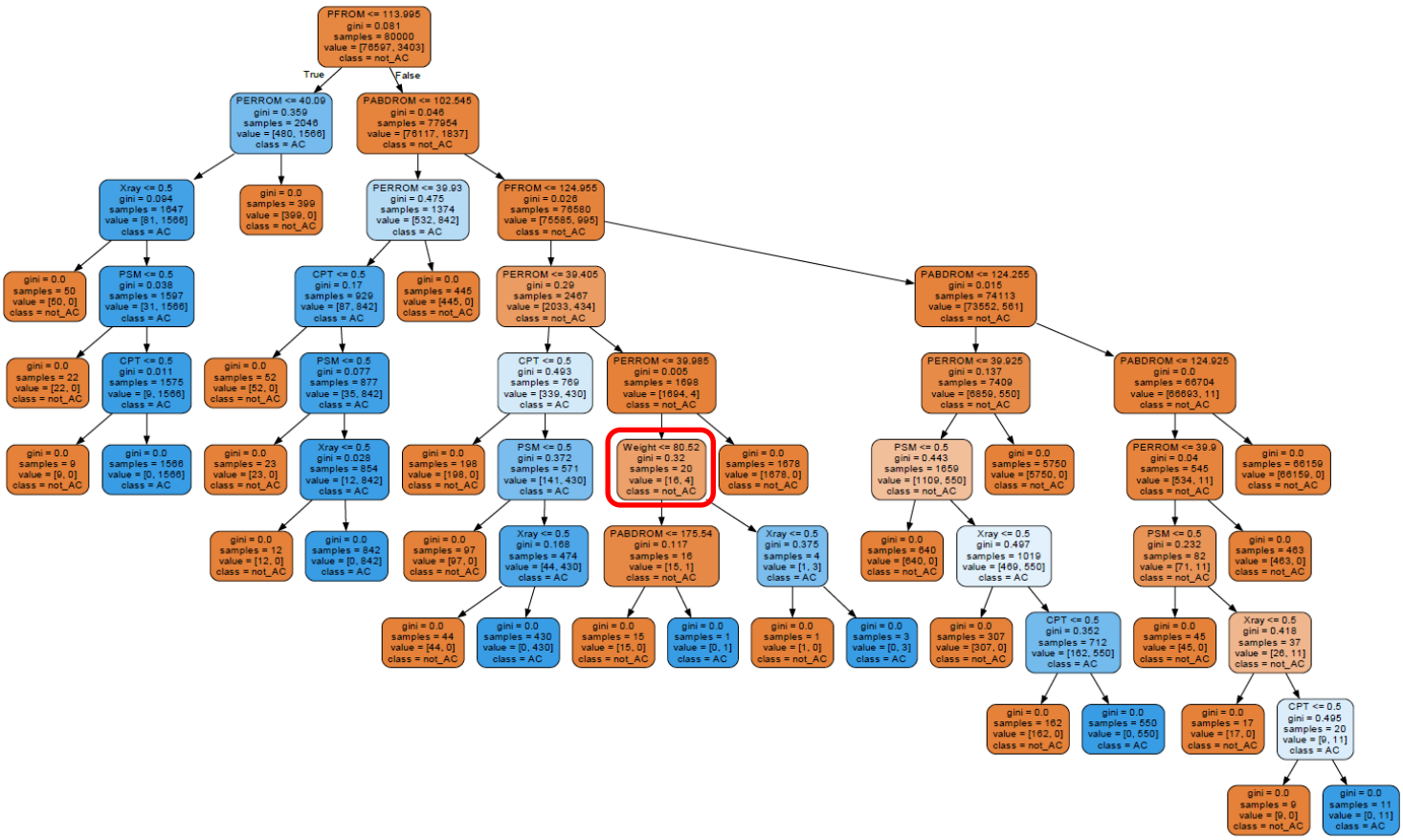
Tree height：9

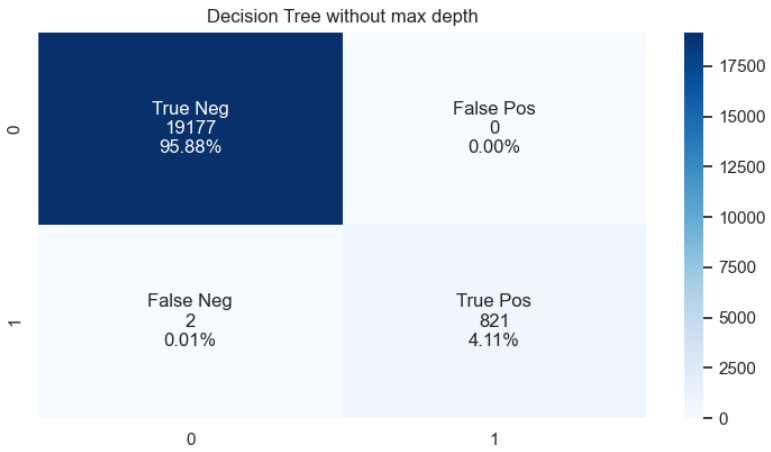
Accuracy：0.999 (2 筆資料分類錯誤)

Precision：1

Recall：0.998

AUC：1





B. data 有 noise 及有標記錯誤(有誤診)

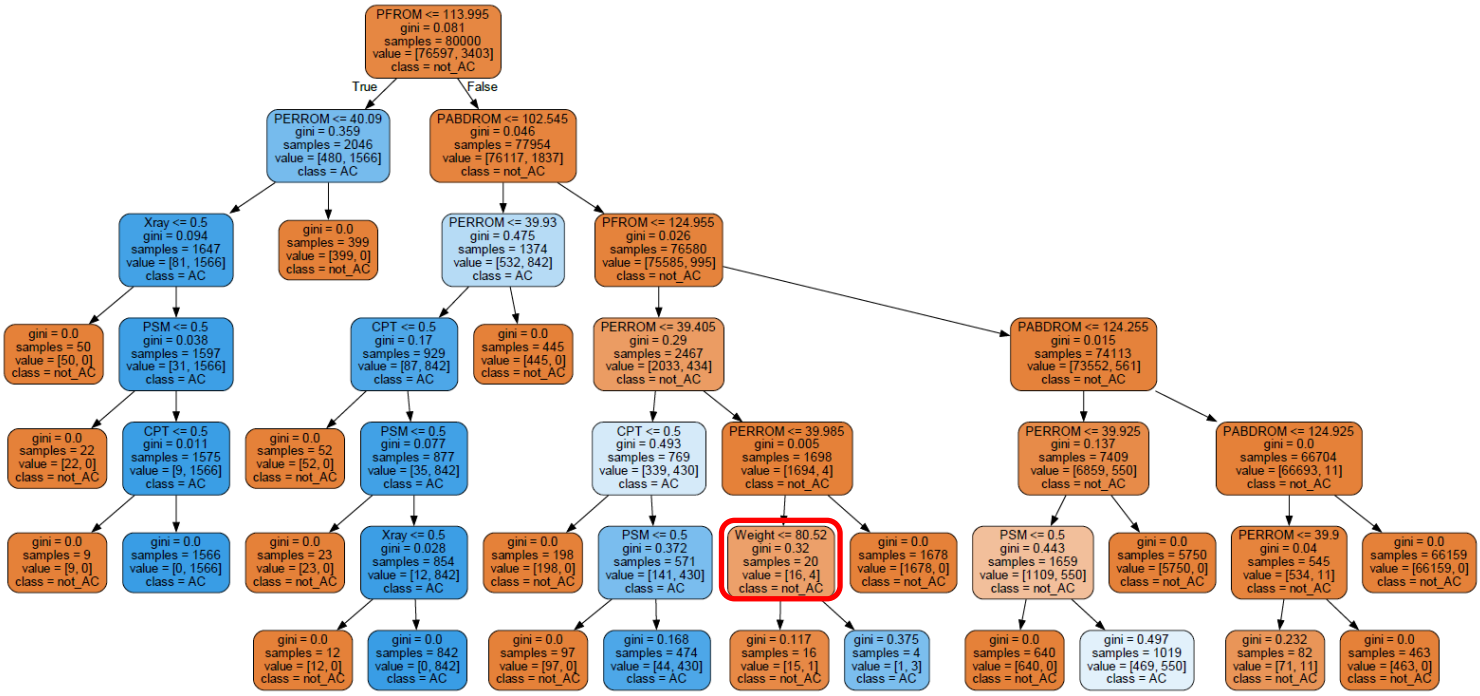
Tree height：6

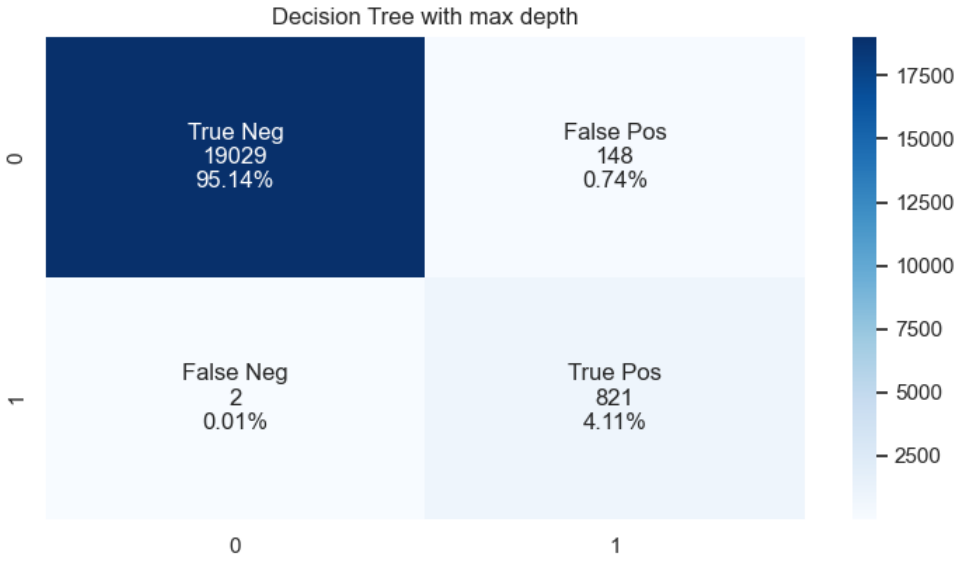
Accuracy：0.993 (150 筆資料分類錯誤)

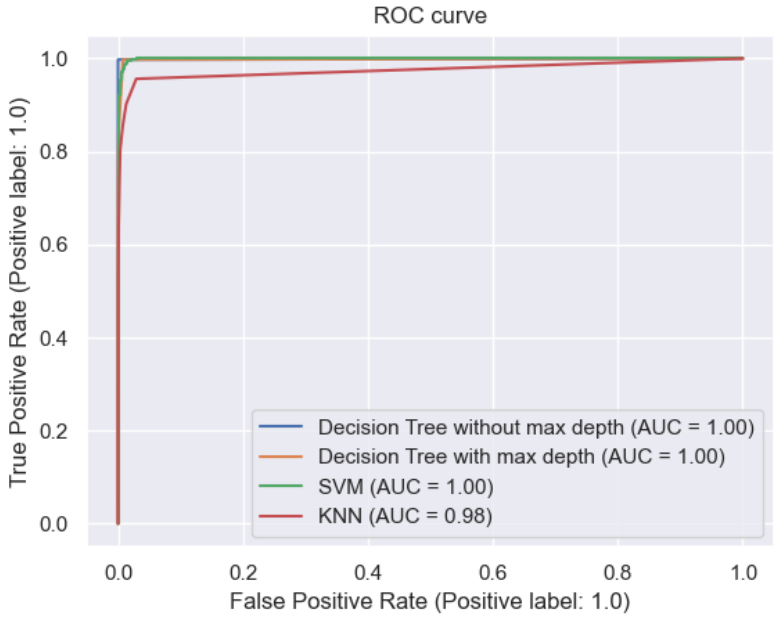
Precision：0.847

Recall：0.998

AUC：1







分析：

在 Dataset 3 中，生成資料有 100000 筆，這次不論有沒有限制樹高的決策樹，找到的幾乎都很接近 Absolutely right rule 的規則進行分類，即使沒找到完整的 Absolutely right rule，找到並使用的分類特徵在實際臨床上也是很重要的判斷標準 (X-ray、PSM、CPT)。

在 A、B 兩個模型中每條路徑和 Dataset 2 一樣也都有找出 PERROM < 40 這個特徵作為 rule 判斷是否患有五十肩。雖然兩個模型都有使用一個較為奇怪的特徵 (Weight) 分類，但只有極為少數的病患有經過這個特徵進行分類，其餘幾乎都是使用正確的規則分類，因此結果我認為相對可以接受。

由於 dataset 2 與 dataset 3 模型所找出的規則都很接近 Absolutely right rule，因此想比較兩個 dataset 訓練出的模型差異。

A 模型在 dataset 3 各方面的表現都較 dataset 2 好，或許是因為 dataset 3 的 data 數量是 dataset 2 的 data 數量 10 倍的關係。但 B 模型在 dataset 2 的 accuracy 與 Precision 表現都較好，應該是被 false positive 數量拖累的關係。但 B 模型在 dataset 3 false negative 的數量較 dataset 2 少，若給予評分的話，由於臨床上 false negative 造成的後果會較 false positive 嚴重 (應該要治療但被診斷 negative 無法及時治療可能使狀況更嚴重)，因此若在給分上面，false negative 會懲罰相對重的分數，因此若要使用在臨床上面的話，dataset 3 的模型表現相對較為可以接受，即使 precision 與 accuracy 較 dataset 2 的模型差一些，但整體表現還是在可接受的範圍內。

從 3 個 dataset 觀察下來，dataset 3 找出來的規則最接近 Absolutely right rule，

且 false negative 的比例最低，因此接下來針對 noise 的分析只會使用 sample = 100000 的 dataset 產生模型，並進行分析與 dataset 3 中的無 noise 的分類結果作為比較。

有誤診(有 noise)的情況

Dataset with noise

Sample 數量：100000 (Training Data : Testing Data = 8:2)

C. 不限制樹高，data 有 noise 及有標記錯誤(有誤診)

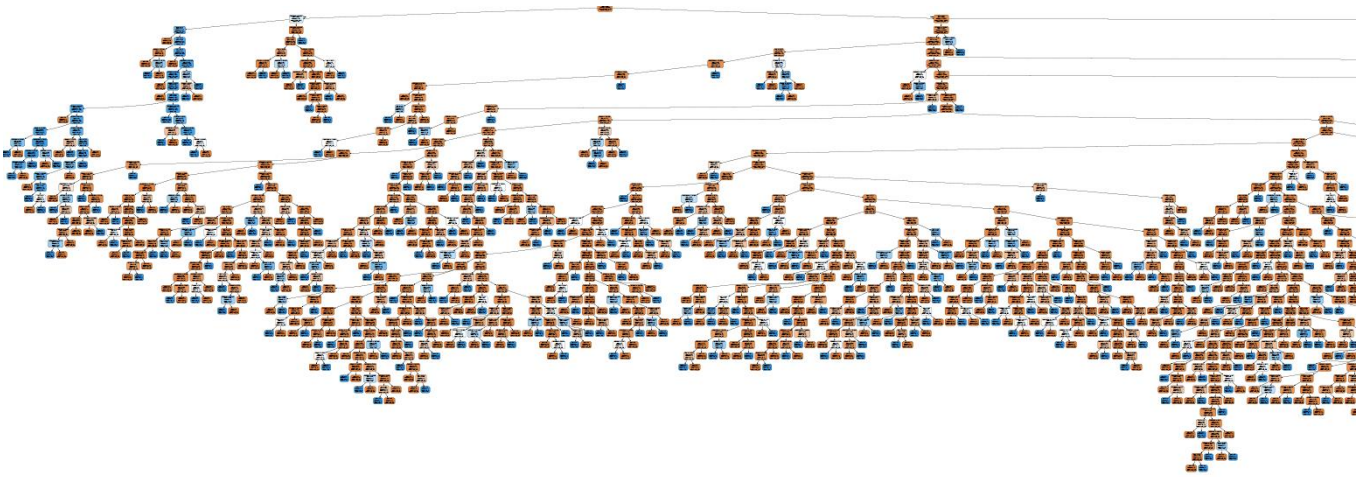
Tree height：非常高，非常深

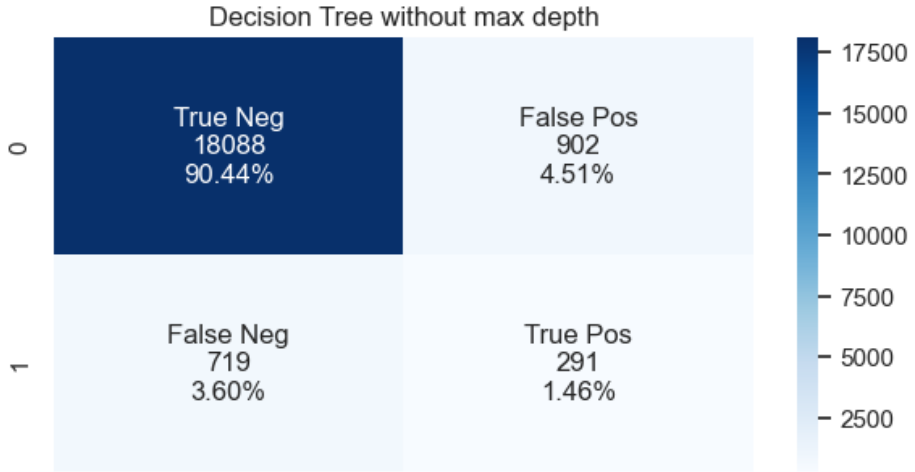
Accuracy：0.919 (1621 筆資料分類錯誤)

Precision：0.244

Recall：0.288

AUC：0.62





D. 限制樹高(max depth = 6)，data 有 noise 及有標記錯誤(有誤診)

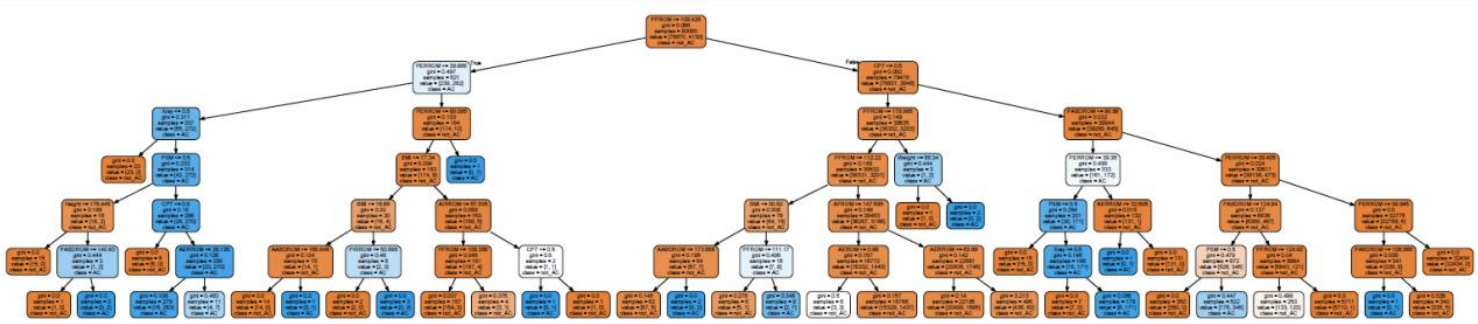
Tree height：6

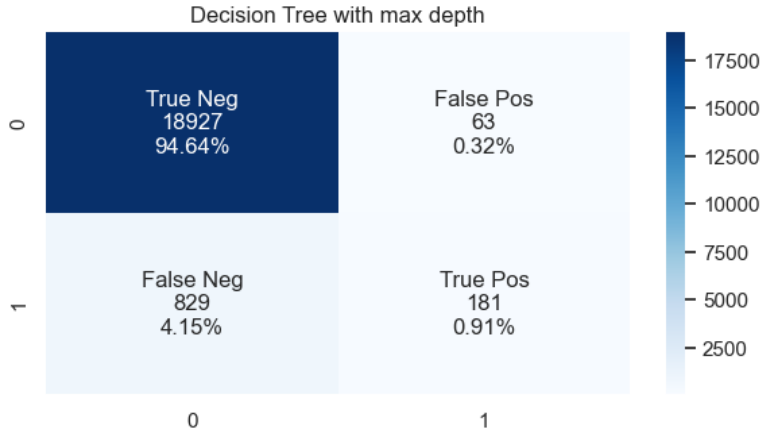
Accuracy：0.955 (892 筆資料分類錯誤)

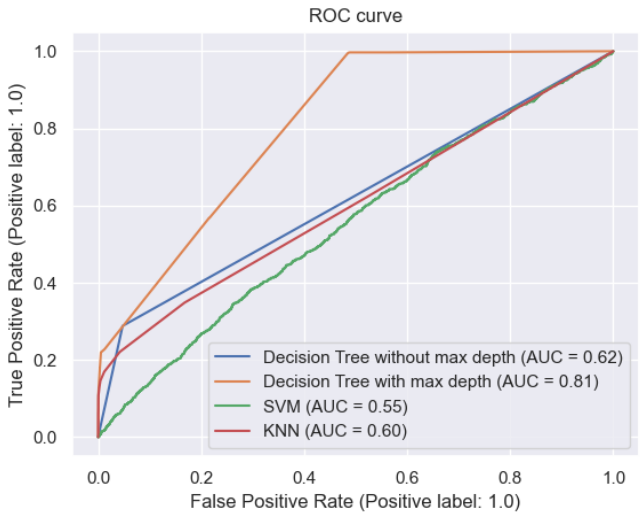
Precision：0.742

Recall：0.179

AUC：0.81







分析：

在 dataset with noise 中，可以看出 C 模型(沒有限制樹高的決策樹)非常寬與深，雖然 Accuracy 有 0.9 以上，不過 Precision 只有 0.244，Recall 只有 0.288，表示分類錯誤的 data 其實不少。另外樹高非常的高，可能有產生 overfitting 的現象，所找到的規則只有很少部分符合 Absolutely right rule，大部分居然會使用到 BMI、Height、與 Weight 去分類，這是非常奇特的現象。

A、C 模型在加入一些 noise 後得到的樹高與樹寬、甚至是規則差距非常大，共通點大概是在 root 點是使用 PFROM 去分類，表示決策樹即便遇到 noise 還是有找出 PFROM 也是一個重要的特徵。

D 模型(max_depth = 6)在 dataset with noise 中，Accuracy = 0.955，Precision = 0.742，不過 Recall 只有不到 0.2，表示應該要被列為 positive 沒有被列為 positive 非常多。首先先查看 D 模型找到的規則，有找到幾條和 Absolutely right rule 接近的路徑，不過有些分類路徑最後也是用 BMI、Height、Weight 進行分類，找出來的規則是錯的，因此雖然 Recall 很低，不過或許是因為有 noise 的關係、實際上真正是五十肩的病患 Recall 應該會再提高一些。和 B 模型相比，加入了一些 noise 的 D 模型分類正確的比例下降不少，表示決策樹對於分類錯誤的 labe 是非常敏感的。

C、D 模型相比，C 模型由於 overfitting 的關係造成 test data accuracy 下降許多，D 模型由於有限制樹高，則沒有產生 overfitting 的狀況，即使有 noise，找出的規則還是有相對接近 absolutely right rule，雖然在沒有 noise 的 dataset 中，A 模型的表現看似相對於 B 模型好，但實際在臨床上不可能完全沒有誤診的情況發生，若要透過決策樹判斷病人是否是五十肩的話，使用 D 模型還是可以得到一定的準確度，且從 AUC 來看，D 模型還有 0.8 以上，有還不錯的鑑別度，反而是 C 模型只有 0.62，似乎和猜的不會差距太多。

結論：

決策樹對於 noise 算是相對敏感的，但在有限制樹高的狀況下，即便有些 Noise 出現，在 sample 量足夠的情況下還是可以得到接近 Absolutely right rule 的分類規則，只是 False negative 的狀況需要注意，recall 數量太低，但在臨床上面會比較在意 false negative，因此若要使用決策樹做為臨床輔助判斷的話或許要再搭配其他的模型一起輔助會較好。

Data analysis with SVM model

Classifier：sklearn.svm.svc (kernel='linear',probability=True)

說明：以下說明將會根據下列兩種情況進行分析，並調整 Sample 的數量觀察是否得到不同的結果。

A. data 無 noise 也沒被標記錯誤(無誤診)
B. data 有 noise 及有標記錯誤(有誤診)

Dataset 1

Sample 數量：1000 (Training Data : Testing Data = 8:2)

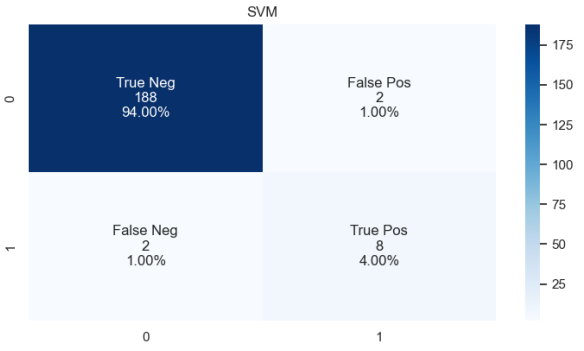
A. data 無 noise 也沒被標記錯誤(無誤診)

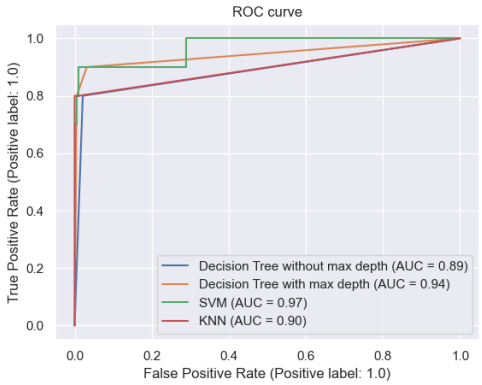
Accuracy：0.98 (4 筆資料分類錯誤)

Precision：0.8

Recall：0.8

AUC：0.97





Dataset 2

Sample 數量：10000 (Training Data : Testing Data = 8:2)

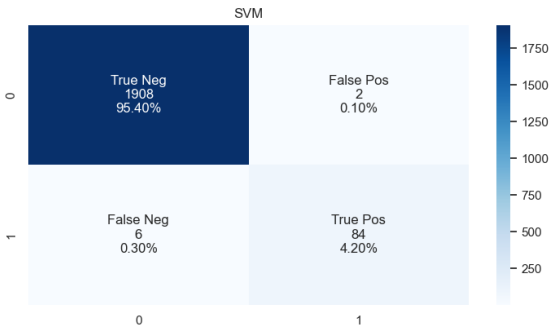
A. data 無 noise 也沒被標記錯誤(無誤診)

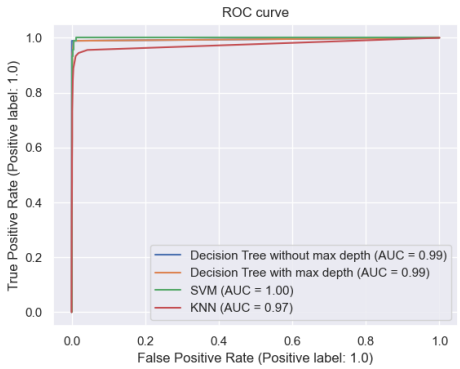
Accuracy：0.96 (8 筆資料分類錯誤)

Precision：0.977

Recall：0.933

AUC：1





Dataset 3

Sample 數量：100000 (Training Data : Testing Data = 8:2)

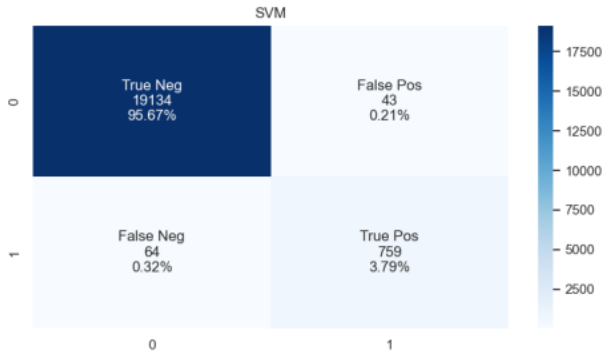
A. data 無 noise 也沒被標記錯誤(無誤診)

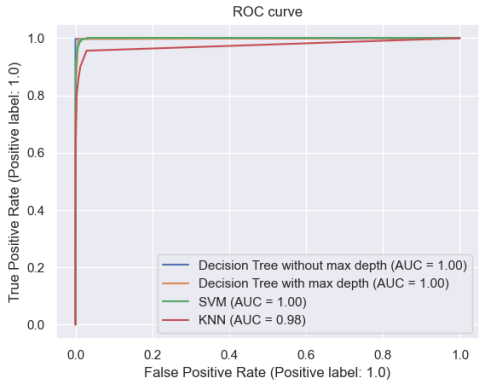
Accuracy：0.995 (107 筆資料分類錯誤)

Precision：0.946

Recall：0.922

AUC：1





Dataset with noise

Sample 數量：100000 (Training Data : Testing Data = 8:2)

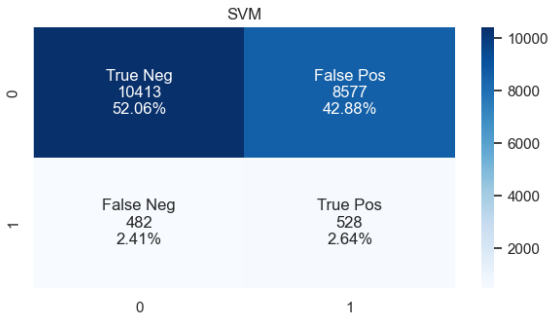
B. data 有 noise 及被標記錯誤(有誤診)

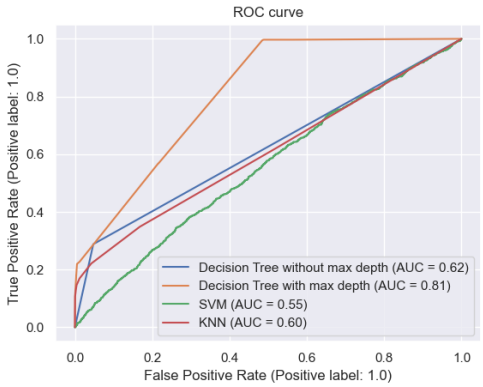
Accuracy：0.547 (9059 筆資料分類錯誤)

Precision：0.058

Recall：0.523

AUC：0.55





分析：

從 Dataset 1~3 可以看出 sample 數量越多，則 SVM 的 accuracy 越高，但可以發現在 false negative 的部分相對於決策樹多了一些，因此雖然 SVM AUC = 1，但臨床上相對不能接受 false negative 的部分，因此以臨床考量來說決策樹會相對較好些。因為在 10 萬筆 sample 的準確度較高，因此測試 dataset with noise 也會以 sample = 100000 進行比較。

原本 SVM 在四種分類器中 AUC 表現是最好的，但在 dataset with noise 中，SVM 反而變成四種模型中分類器中 AUC 表現最差，可能的原因推測是因為 SVM 是找出一個 hyperplane 使資料可以分成兩堆，但使用資料的 Attribute 中有 random 產生的數值，而且又有一些 label 錯誤的的狀況，導致 SVM 很難將資料漂亮地分成兩堆。原本 data 較乾淨時(無誤診)可以漂亮地分成兩堆，因此準確度與 AUC 非常高。

另外 SVM 在尋找 hyperplane 時，在 sample > 10000 筆後分析時間大幅增加才能跑出結果，加上這次 data 中又含有 noise，沒有設定 iteration 限制時是跑不出結果的，後來將 max_iteration 設定為 1000 才順利跑出結果，或許是 iteration 次數還不夠多的原因，也早成 dataset with noise 結果不太好。AUC = 0.55 幾乎等於模型是用猜的，完全沒辦法作為什麼參考。

目前想到可以使模型進步的方式是，若是先使用降維，先將一些共線性高的特徵移除(ROM 系列)，只留下獨立的特徵讓維度下降，或許跑出來的結果可以好一些，或是使用其他的 kernel 作分類。

Data analysis with KNN model

Classifier： sklearn.neighbors.KNeighborsClassifier

說明：以下說明將會根據下列兩種情況進行分析，並調整 Sample 的數量觀察是否得到不同的結果且 KNN 會使用的 K 值會從 K=1~20 中找出 error rate 最小的 K 來訓練模型。

A. data 無 noise 也沒被標記錯誤(無誤診)
B. data 有 noise 及有標記錯誤(有誤診)

Dataset 1

Sample 數量：1000 (Training Data : Testing Data = 8:2)

A. data 無 noise 也沒被標記錯誤(無誤診)

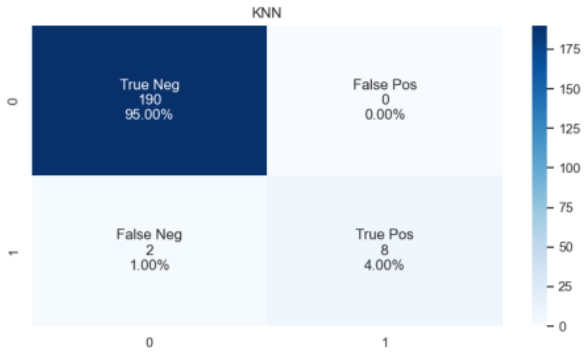
Accuracy：0.99 (2 筆資料分類錯誤)

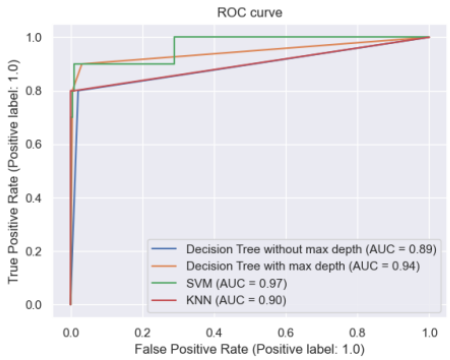
Precision：1

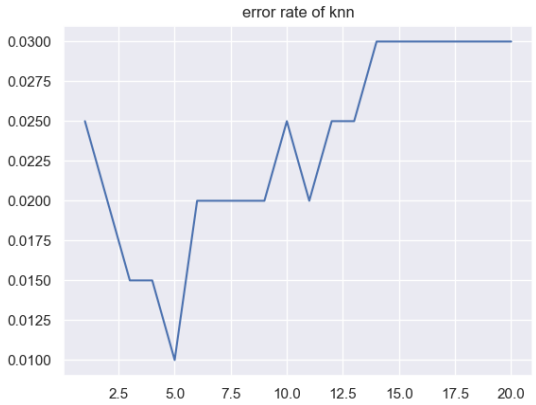
Recall：0.8

AUC：0.9

K = 5







Dataset 2

Sample 數量：10000 (Training Data : Testing Data = 8:2)

A. data 無 noise 也沒被標記錯誤(無誤診)

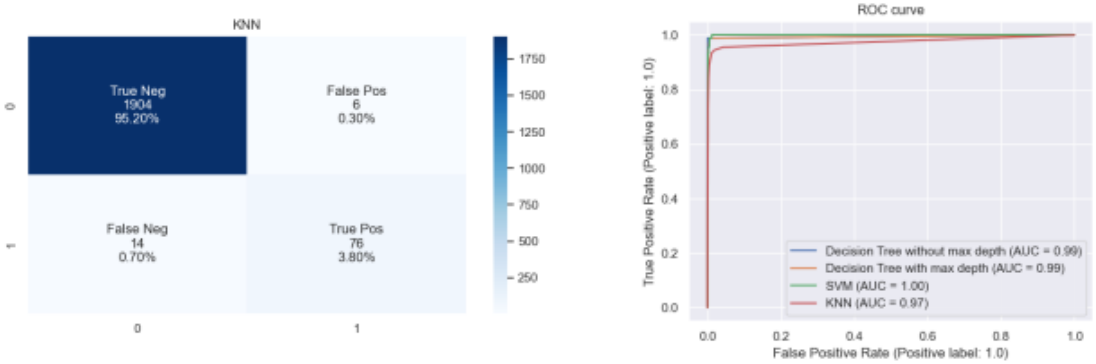
Accuracy：0.99 (20 筆資料分類錯誤)

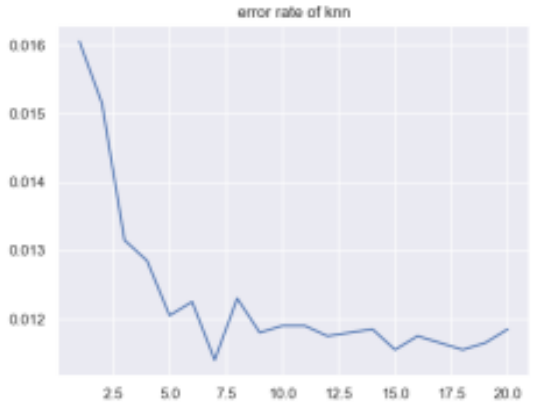
Precision：0.927

Recall：0.844

AUC：0.97

K = 7





Dataset 3

Sample 數量：100000 (Training Data : Testing Data = 8:2)

A. data 無 noise 也沒被標記錯誤(無誤診)

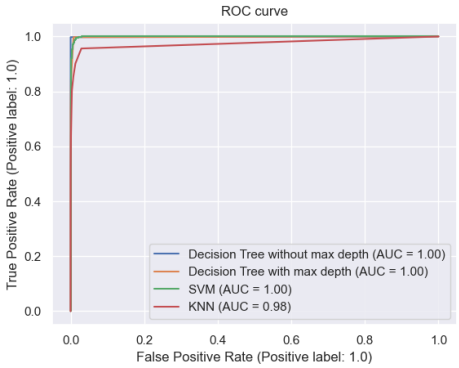
Accuracy：0.987 (228 筆資料分類錯誤)

Precision：0.908

Recall：0.804

AUC：0.98

K = 7



Dataset with noise

Sample 數量：100000 (Training Data : Testing Data = 8:2)

B. data 有 noise 及被標記錯誤(有誤診)

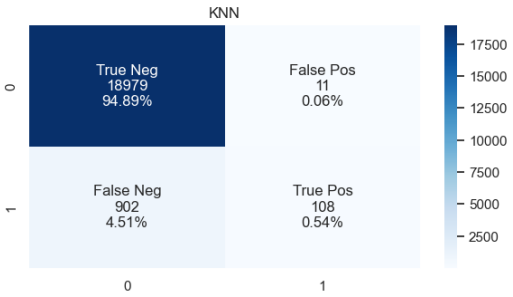
Accuracy：0.954 (913 筆資料分類錯誤)

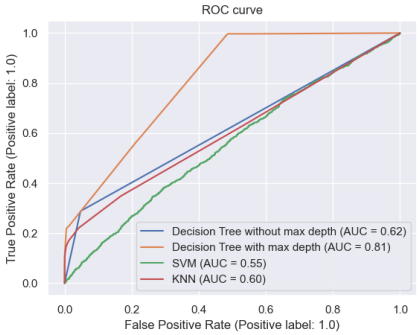
Precision：0.908

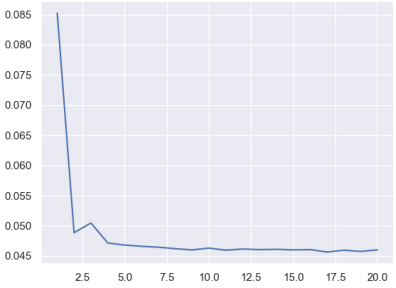
Recall：0.107

AUC：0.6

K = 17







分析：

從 Dataset 1~3 可以看出 sample 數量越多，則 KNN 的 accuracy 越高，但可以發現在 false negative 的部分相對其他模型多非常多，因此雖然 KNN AUC > 0.95 且 Accruacy > 0.95，但臨床上相對不能接受 false negative 的部分，因此以臨床考量來說決策樹會相對較好些。因為在 10 萬筆 sample 的準確度較高，因此測試 dataset with noise 也會以 sample = 100000 進行比較。

在 dataset with noise 中，KNN 雖然 accuracy 有達到 0.95 以上，不過 Recall 的部分非常低，加入 noise 後 false negative 非常多，而且 AUC 只剩下 0.6，基本上和猜的程度差不多，推測可能是因為加入的 Noise 算是隨機的，而非有線性關係，讓 KNN 難以辨別被標記的 label 哪個才是正確的。

雖然 KNN 有一定的 accuracy，但從 AUC 來看 decision tree 對於有 noise 的 data 表現還是較好。

結論：

三個模型中，若要套用到臨床作為輔助判斷五十肩的模型，決策樹看起來會是會佳的選擇，即便加入 noise，AUC 也還有 0.8 以上，若要再多使用一個模型作為輔助判斷，KNN 和決策樹的組合會是比較好的，但如果測出來病患被歸類在 negative 的話，可能要依決策樹的判斷為主比較不會有錯誤，畢竟 KNN 在 false negative 上比例不低。

本次實驗模擬臨床上五十肩可能會作的理學檢查，其中 AROM 與 PROM 會是有高度相關性的關係，因此決策樹有時會將 AROM 作為特徵判斷也算是相對可以接受，但這次跑出來的模型，基本上決策樹還是有抓到說以 PROM 作為特徵判斷，其餘重要的理學檢查特徵也有準確的找出來，不過想想其實也蠻合理的，理學檢查中許多 diagnosis 都是 if 符合某個條件，再多測試另一個條件，逐步逼近自己對於疾病的推測，這和決策樹的核心概念其實很接近，因此決策樹會是表現最好的模型也不是很意外。只是當 noise 增加時，由於決策樹對於 Noise 高度敏感，若 noise 過多的話整顆樹可能就會被摧毀，因此誤診 data 也不能太多，或是整體收集的 sample 要夠多才能稀釋掉錯誤 label 的 data 帶來的影響。