資料探勘期末

子宮頸癌風險預測

人工智慧四 10812243 陳思維

資料集介紹 ：Kaggle 上的 UCI Machine learing :

Cervical Cancer Risk Classification

https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification

資料概況: 36 columns x 859 rows

我們的目的是要藉由 Dataset 提供各種的不良生活習慣，多項致癌風險因子、不同的身體疾病和各項檢驗結果等等，去預測此人是否得到或有高機率罹患子宮頸癌。

資料欄位介紹

Column	代表意思	Type
Age	年齡	整數連續值
Number of sexual partners	性伴侶數	整數連續值
First sexual intercourse	第一次性經驗年齡	整數連續值
Num of pregnancies	懷孕次數	整數連續值
Smokes	有無抽菸	0 : No Smoke 1 : Smoke
Smokes (years)	抽菸時間 (年)	非整數連續值
Smokes (packs/year)	一年抽幾包	非整數連續值
Hormonal Contraceptives	是否使用過賀爾蒙避孕藥	0 : No 1 :Yes
Hormonal Contraceptives (years)	賀爾蒙避孕藥使用年數	非整數連續值
IUD	是否使用宮內結孕器	0 : No 1 :Yes
IUD (years)	子宮內結孕器使用年數	非整數連續值

資料欄位介紹

STDs (number)	性病數量	整數連續值
Dx : HPV	有無感染人類乳突病毒	布林值
Dx : CIN	有無**子宮頸上皮內瘤變**	布林值
目標變量 : Dx:Cancer	是否已確診子宮頸癌	布林值
目標變量 : Hinselmann	陰道鏡檢查是否異常	布林值
目標變量 : Schiller	是否有生殖性細胞腫瘤	布林值
目標變量 : Citology	細胞檢驗是否異常	布林值
目標變量 : Biopsy	切片檢驗是否異常	布林值
目標變量 : Cancer	上述四個目標變量，有兩個異常便可	以確診子宮頸癌

範例頁尾文字 20XX 年 <編號>

我們的目標變量是 Cancer，Dx:Cancer 是已確診子宮頸癌

Cancer 此欄位是依照目標變量 : Hinselmann、 Schiller、Citology、Biopsy

這四項檢驗結果去評估，只要兩項檢驗異常，即可確診子宮頸癌，另外，

Dx : Cancer (已經確診子宮頸癌)這個目標變量，只要布林值為 1，Cancer 值直接填入 1。

一、資料前處理 :

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)	STDs:condylomatosis	STDs:cervi condylomat
0	18	4	15	1	0	0	0	0	0	0	0	0	0	0	
1	15	1	14	1	0	0	0	0	0	0	0	0	0	0	
2	34	1	NaN	1	0	0	0	0	0	0	0	0	0	0	
3	52	5	16	4	1	37	37	1	3	0	0	0	0	0	
4	46	3	21	4	0	0	0	1	15	0	0	0	0	0	

這是 dataframe 的形式，先查看每個欄位的資料型態

Age	int64
Number of sexual partners	object
First sexual intercourse	object
Num of pregnancies	object
Smokes	object
Smokes (years)	object
Smokes (packs/year)	object
Hormonal Contraceptives	object
Hormonal Contraceptives (years)	object
IUD	object
IUD (years)	object
STDs	object
STDs (number)	object
STDs:condylomatosis	object
STDs:cervical condylomatosis	object
STDs:vaginal condylomatosis	object
STDs:vulvo-perineal condylomatosis	object
STDs:syphilis	object
STDs:pelvic inflammatory disease	object
STDs:genital herpes	object
STDs:molluscum contagiosum	object
STDs:AIDS	object
STDs:HIV	object
STDs:Hepatitis B	object
STDs:HPV	object

Age	int64
Number of sexual partners	float64
First sexual intercourse	float64
Num of pregnancies	float64
Smokes	float64
Smokes (years)	float64
Smokes (packs/year)	float64
Hormonal Contraceptives	float64
Hormonal Contraceptives (years)	float64
IUD	float64
IUD (years)	float64
STDs	float64
STDs (number)	float64
STDs:condylomatosis	float64
STDs:cervical condylomatosis	float64
STDs:vaginal condylomatosis	float64
STDs:vulvo-perineal condylomatosis	float64
STDs:syphilis	float64
STDs:pelvic inflammatory disease	float64
STDs:genital herpes	float64
STDs:molluscum contagiosum	float64
STDs:AIDS	float64
STDs:HIV	float64
STDs:Hepatitis B	float64
STDs:HPV	float64

Step1： 將 object 型態轉為 float 型態

Step2： 查看 kaggle 上的欄位統計圖，刪除空缺值過多欄位。

Step3： 填補空缺值，若欄位是連續值，補平均值

若是布林值則觀察欄位填補 0 or 1 值。

```
def convert_median(name0 = '') :
    FirstData[name0] = FirstData[name0].fillna(FirstData[name0].median())           # 補缺失值: 平均

convert_median('Number of sexual partners')
convert_median('First sexual intercourse')
convert_median('Num of pregnancies')
convert_median('Smokes (years)')
convert_median('Smokes (packs/year)')
convert_median('Hormonal Contraceptives (years)')
convert_median('STDs (number)')
convert_median('STDs:condylomatosis')
convert_median('STDs:vulvo-perineal condylomatosis')
convert_median('STDs:syphilis')
convert_median('STDs:HIV')

def convert_0_1( name1 = '', num1 = 0 ) :
    FirstData[name1] = FirstData[name1].fillna(0)          # 補缺失值: 補0 or 1

convert_0_1('IUD', 0)
convert_0_1('IUD (years)', 0)
convert_0_1('Smokes', 1)
convert_0_1('STDs', 1)
convert_0_1('Hormonal Contraceptives', 1)
```

Step4 : Hinselmann、 Schiller、Citology、Biopsy 這四項癌症檢驗目標函數

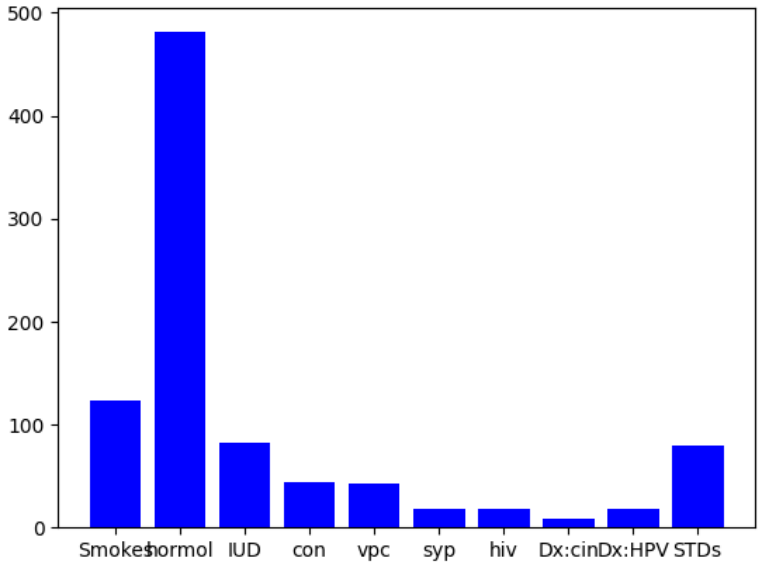
符合兩項，或者 Dx: Cancer 為 1 則確診子宮頸癌，在 Cancer 欄位填入 1

```
for i in range(len(FirstData)) :
    if FirstData['Hinselmann'][i] + FirstData['Schiller'][i] + FirstData['Citology'][i] + FirstData['Biopsy'][i] > 1 :
        FirstData['Cancer'][i] = 1

    else :                                                          # 4 種目標函數，只要符合兩種檢查欄位是陽性的，便是確診，Cancer是我自己多弄出的欄位
        FirstData['Cancer'][i] = FirstData['Dx:Cancer'][i]

    i = i + 1
```

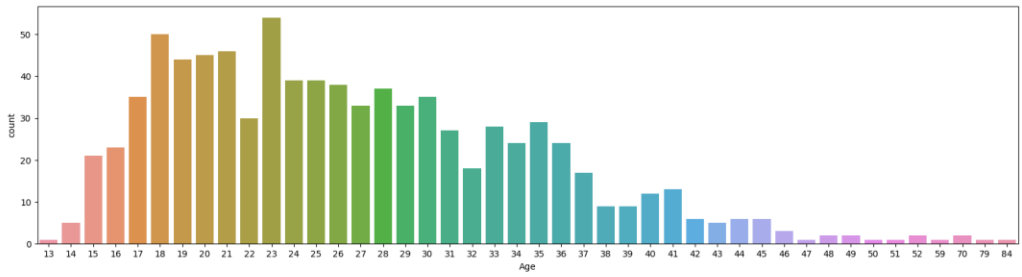
二、特徵統計及視覺化

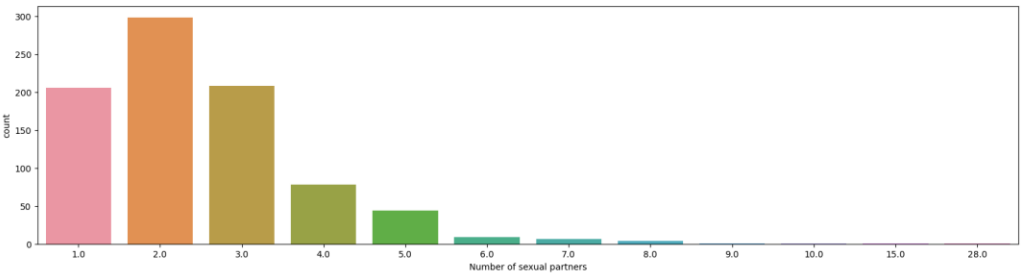
1. 計算離散型風險因子陽性數量

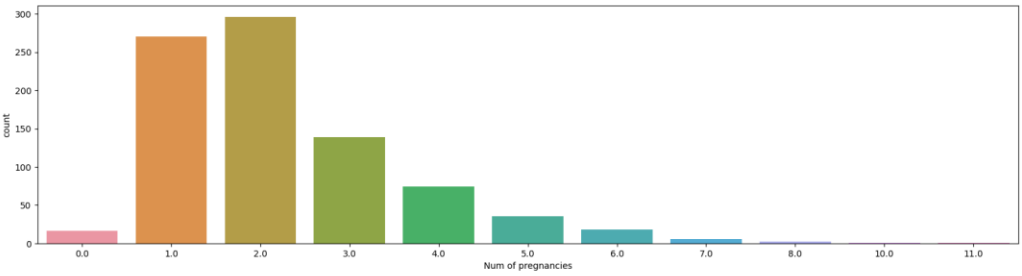
```
Smokes :  123.0
Hormonal Contraceptives :  481.0
IUD :  83.0
STDs:condylomatosis :  44.0
STDs:vulvo-perineal condylomatosis :  43.0
STDs:syphilis :  18.0
STDs:HIV :  18.0
Dx:CIN :  9.0
Dx:HPV :  18.0
STDs :  79.0
```

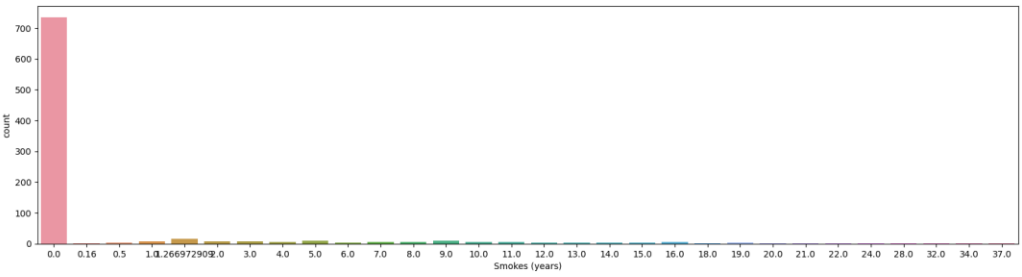


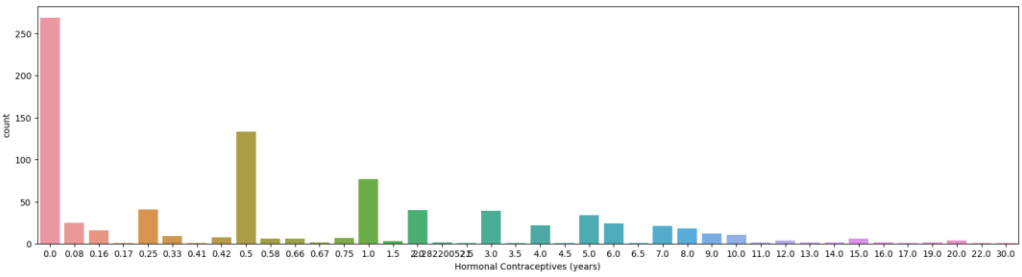
2. 連續數值風險因子欄位圖表

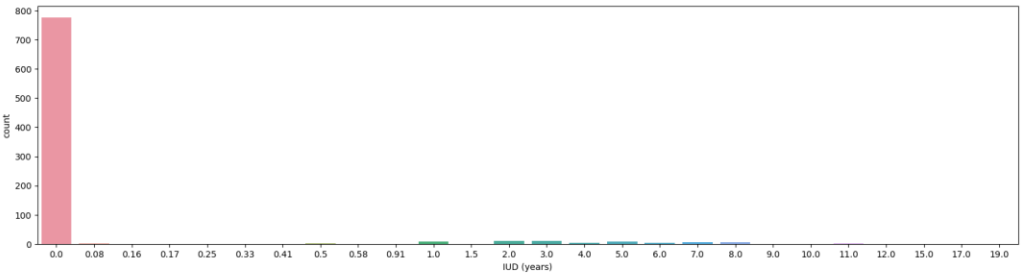

count
Age

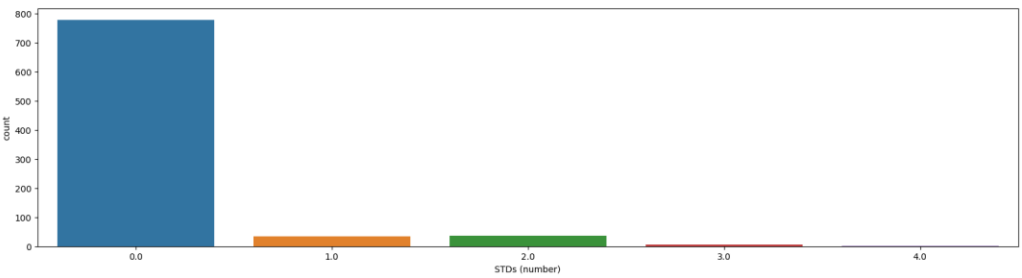

count
Number of sexual partners


count
Num of pregnancies

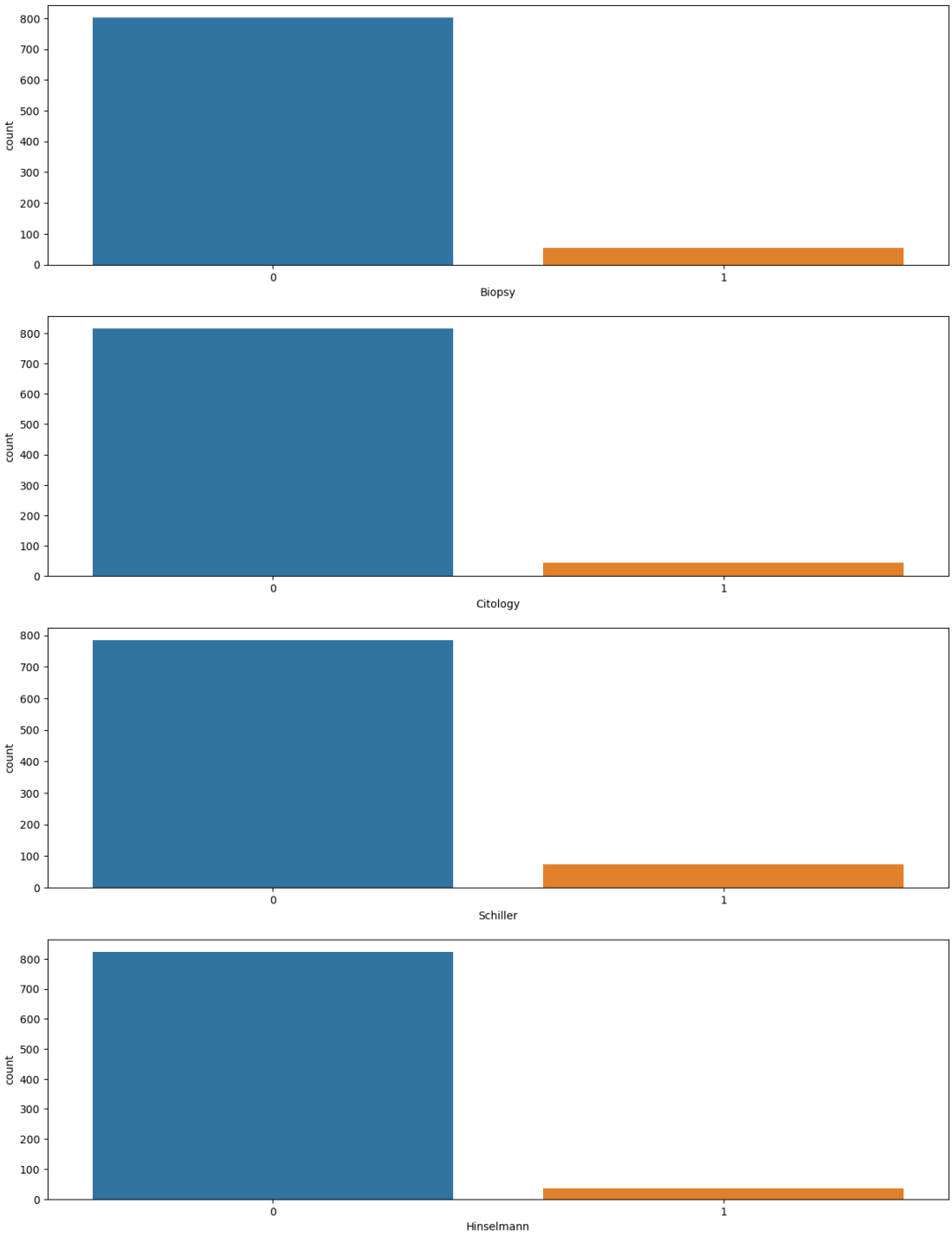

count
Smokes (years)


count
Hormonal Contraceptives (years)

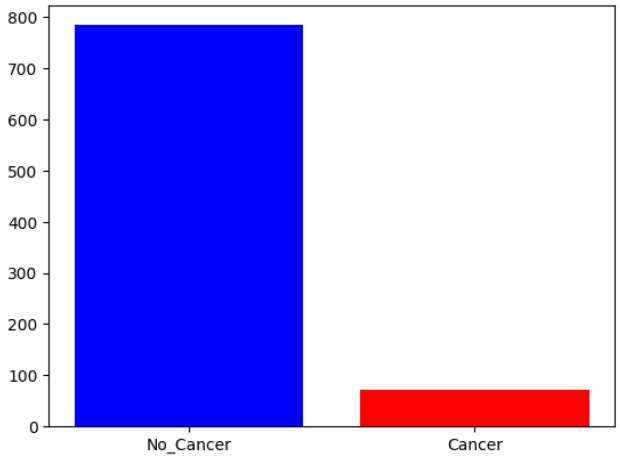

count
IUD (years)


count
STDs (number)

3. 癌症檢查結果統計



4. 是否罹患癌症統計



三、訓練與測試以及模型評估

```
X_train, X_test, Y_train, Y_test = train_test_split(train_x, train_y,
test_size = 0.25, random_state = 12)  # 劃分訓練集，測試集
```

1. 將訓練集及測試集比例分成 80%及 20，random_state 設定為 12
2. 使用 Decision tree Model ，根據 Entropy、Gini、leaf_nodes、depth 係數分成 4 種參數組合。

```
model0 = DecisionTreeClassifier(criterion="entropy", max_leaf_nodes =
2)   # Entropy node Decision tree
model0.fit(X_train, Y_train)
yyyyy = model0.predict(X_test)

model = DecisionTreeClassifier(criterion="entropy", max_depth= 3)   #
Entropy depth Decision tree
model.fit(X_train, Y_train)
yyyyy = model.predict(X_test)

def gini_node_tree(num):
    model = DecisionTreeClassifier(criterion="gini", max_leaf_nodes =
num)   # Gini node Decision tree
    model.fit(X_train, Y_train)
def gini_depth_tree(num):
    modelx = DecisionTreeClassifier(criterion="gini", max_depth= num)  #
    Gini depth Decision tree
    modelx.fit(X_train, Y_train)
    yyyyy = modelx.predict(X_test)
```

3. 決策樹模型最大準確率及 Recall & F1score

```
正確率為 0.9534883720930233
recall: 0.9
f1_score: 0.6428571428571429
```

4. 使用隨機森林的最大準確率及 Recall & F1score

```
modely = RandomForestClassifier(criterion="entropy", max_leaf_nodes=
5)  #  隨機森林
modely.fit(X_train, Y_train)
yyyyy = modely.predict(X_test)
recall = recall_score(yyyyy,Y_test)
f1_score = f1_score(yyyyy,Y_test)
```

```
正確率為 0.9209302325581395
recall: 1.0
f1_score: 0.10526315789473684
```

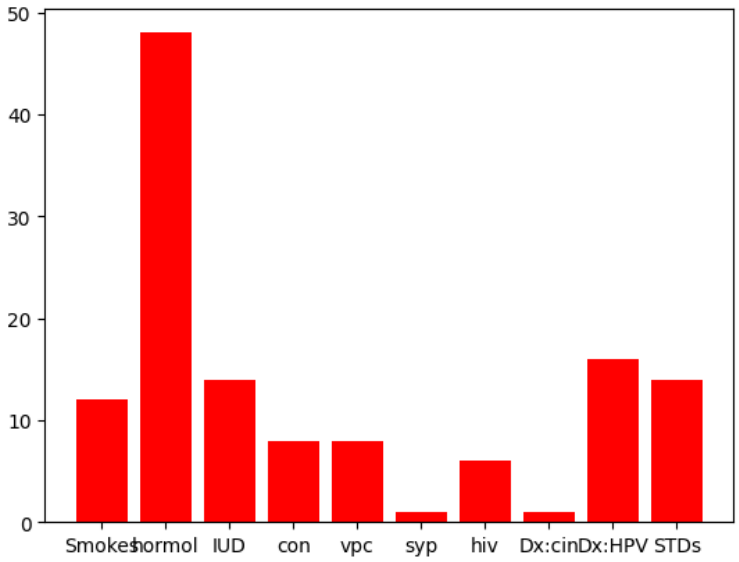
四、評估影響風險因子較大的特徵

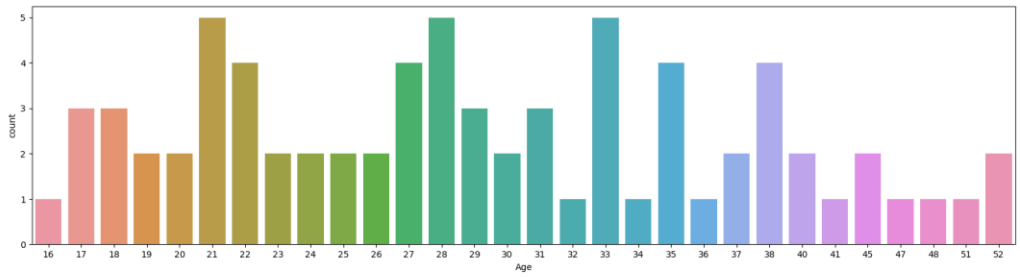
1. 利用 eli5 套件分析與計算欄位權重值

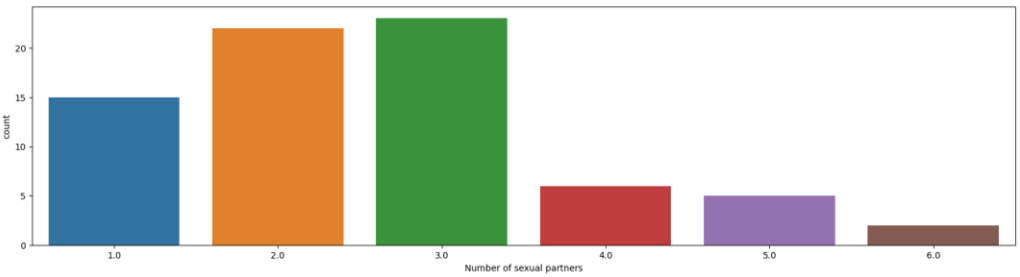
權重	特徵
0.0614 +- 0.0199	Dx : HPV
0.0140 +- 0.0156	Num of preganancies
0.0084 +- 0.0108	Hormonal Contraceptives (years)
0.0056 +- 0.0037	Age
0.0047 +- 0.0000	Number of sexual partners
0.0037 +- 0.0123	Hormonal Contraceptives
0.0028 +- 0.0046	First sexual intercourse
0.0009 +- 0.0037	STDs

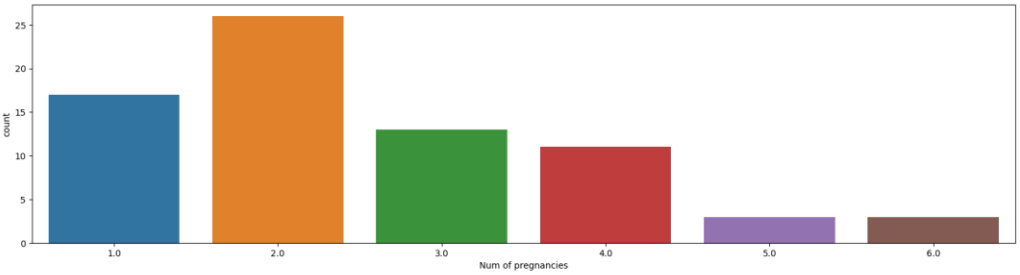
2. 罹患癌症的人各項風險因子統計

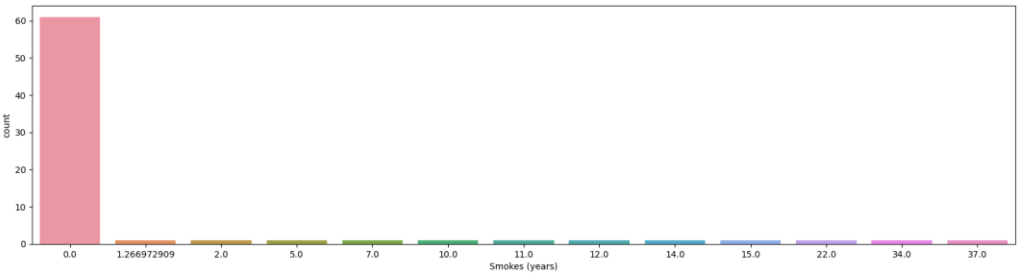
```
Cancer :  73.0
Smokes :  12.0
Hormonal Contraceptives :  48.0
IUD :  14.0
STDs:condylomatosis :  8.0
STDs:vulvo-perineal condylomatosis :  8.0
STDs:syphilis :  1.0
STDs:HIV :  6.0
Dx:CIN :  1.0
Dx:HPV :  16.0
STDs :  14.0
```

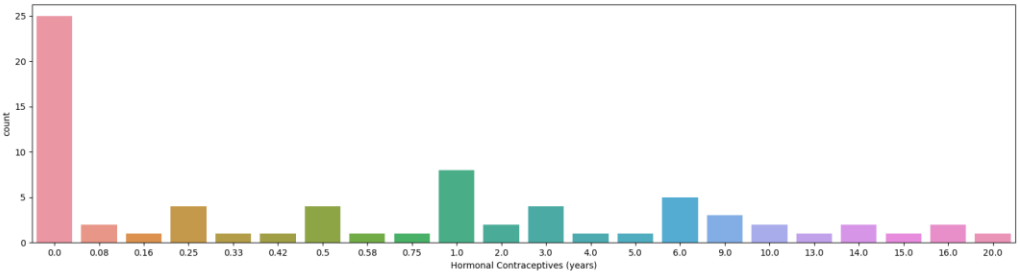


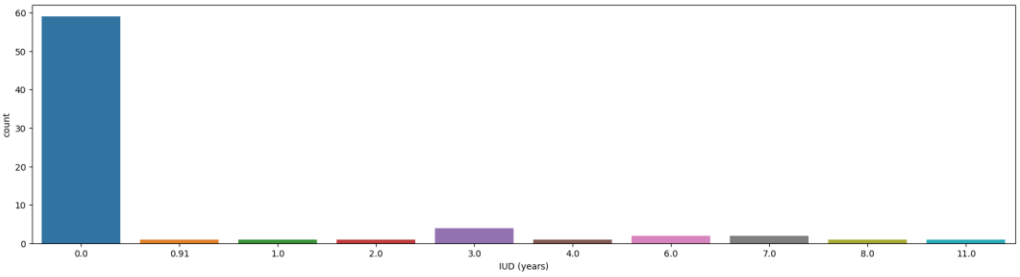

count
0 1 2 3 4 5
16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 40 41 45 47 48 51 52
Age

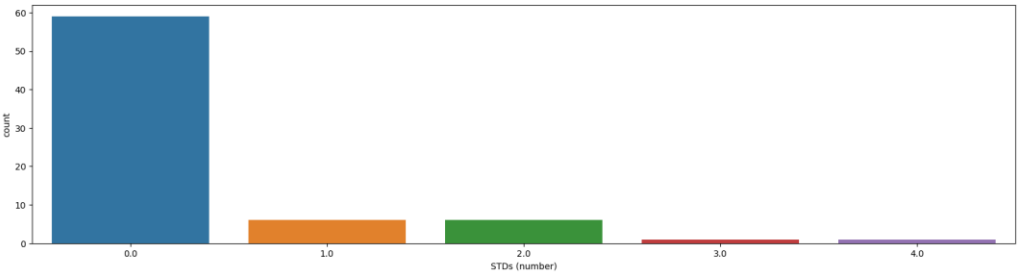

count
0 5 10 15 20
1.0 2.0 3.0 4.0 5.0 6.0
Number of sexual partners


count
0 5 10 15 20 25
1.0 2.0 3.0 4.0 5.0 6.0
Num of pregnancies

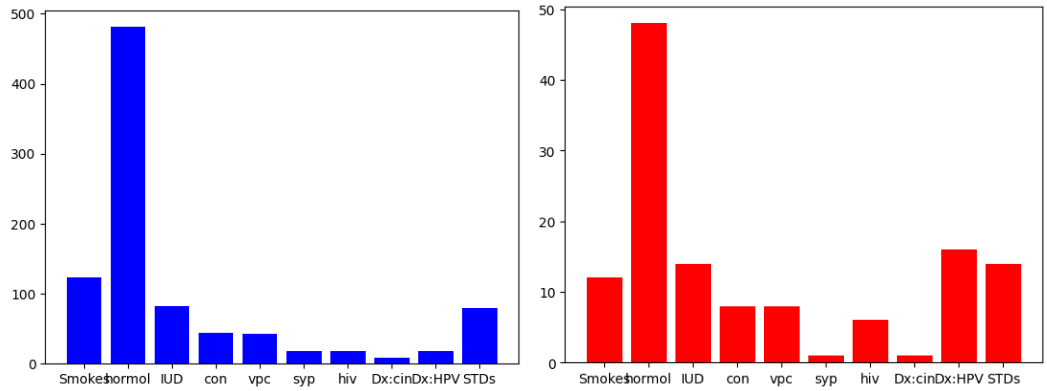

count
0 10 20 30 40 50 60
0.0 1.266972909 2.0 5.0 7.0 10.0 11.0 12.0 14.0 15.0 22.0 34.0 37.0
Smokes (years)


count
0 5 10 15 20 25
0.0 0.08 0.16 0.25 0.33 0.42 0.5 0.58 0.75 1.0 2.0 3.0 4.0 5.0 6.0 9.0 10.0 13.0 14.0 15.0 16.0 20.0
Hormonal Contraceptives (years)


count
0 10 20 30 40 50 60
0.0 0.91 1.0 2.0 3.0 4.0 6.0 7.0 8.0 11.0
IUD (years)


count
0 10 20 30 40 50 60
0.0 1.0 2.0 3.0 4.0
STDs (number)

五、結果與討論



從上兩張圖以及權重結果分析來看，可以找出風險因子最大的是 Dx:HPV，測試人有罹患 Dx:HPV 有 18 人，而確診子宮頸癌的人有 16 人有 Dx:HPV，再來第二高及第三高風險因子分別是 Num of preganancies 以及 Hormonal Contraceptives (years)。