P~S 엔트로피 조사

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
import math
```

In [2]:
```
import os
# 운영체제별 한글 폰트 설정
if os.name == 'posix': # Mac 환경 폰트 설정
    plt.rc('font', family='AppleGothic')
elif os.name == 'nt': # Windows 환경 폰트 설정
    plt.rc('font', family='Malgun Gothic')

plt.rc('axes', unicode_minus=False) # 마이너스 폰트 설정


# 글씨 선명하게 출력하는 설정
%config InlineBackend.figure_format = 'retina'
```

In [3]:
```
data1 = pd.read_csv('NHIS_OPEN_GJ_2017_100.csv',encoding='euc-kr')
data2 = pd.read_csv('NHIS_OPEN_GJ_2018_100.csv')
```

In [4]:
```
data1.columns
```

Out[4]:
```
Index(['기준년도', '가입자일련번호', '성별코드', '연령대코드(5세단위)', '시도코드', '신장(5Cm단위)',
       '체중(5Kg단위)', '허리둘레', '시력(좌)', '시력(우)', '청력(좌)', '청력(우)', '수축기혈압',
       '이완기혈압', '식전혈당(공복혈당)', '총콜레스테롤', '트리글리세라이드', 'HDL콜레스테롤', 'LDL콜레스테롤',
       '혈색소', '요단백', '혈청크레아티닌', '(혈청지오티)AST', '(혈청지오티)ALT', '감마지티피', '흡연상태',
       '음주여부', '구강검진수검여부', '치아우식증유무', '결손치유무', '치아마모증유무', '제3대구치(사랑니)이상', '치석',
       '데이터공개일자'],
      dtype='object')
```

In [5]:
```
data1 = data1[['식전혈당(공복혈당)','총콜레스테롤', '트리글리세라이드', 'HDL콜레스테롤', 'LDL
data2 = data2[['식전혈당(공복혈당)','총콜레스테롤', '트리글리세라이드', 'HDL콜레스테롤', 'LDL
```

당뇨병 엔트로피 구하기

2017, 2018년도 데이터 합치기

In [6]:
```
data = pd.concat([data1, data2])
data
```

Out[6]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
0	99.0	193.0	92.0	48.0	126.0
1	106.0	228.0	121.0	55.0	148.0
2	98.0	136.0	104.0	41.0	74.0
3	95.0	201.0	106.0	76.0	104.0
4	101.0	199.0	104.0	61.0	117.0
...	...	...	...	...	...
999995	107.0	NaN	NaN	NaN	NaN
999996	114.0	NaN	NaN	NaN	NaN
999997	98.0	NaN	NaN	NaN	NaN
999998	94.0	NaN	NaN	NaN	NaN
999999	85.0	NaN	NaN	NaN	NaN

2000000 rows × 5 columns

null인 데이터 제거

In [7]:

```
data = data.dropna(axis = 0)
data
```

Out[7]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
0	99.0	193.0	92.0	48.0	126.0
1	106.0	228.0	121.0	55.0	148.0
2	98.0	136.0	104.0	41.0	74.0
3	95.0	201.0	106.0	76.0	104.0
4	101.0	199.0	104.0	61.0	117.0
...	...	...	...	...	...
999978	92.0	158.0	139.0	70.0	60.0
999979	82.0	274.0	239.0	45.0	181.0
999981	99.0	216.0	122.0	43.0	148.0
999982	95.0	222.0	173.0	39.0	148.0
999983	105.0	148.0	214.0	42.0	63.0

1322951 rows × 5 columns

혈당 수치가 200 이상인 데이터 제거

In [8]:

```
data = data[data['식전혈당(공복혈당)'] < 200]
data
```

Out[8]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
0	99.0	193.0	92.0	48.0	126.0
1	106.0	228.0	121.0	55.0	148.0
2	98.0	136.0	104.0	41.0	74.0
3	95.0	201.0	106.0	76.0	104.0
4	101.0	199.0	104.0	61.0	117.0
...	...	...	...	...	...
999978	92.0	158.0	139.0	70.0	60.0
999979	82.0	274.0	239.0	45.0	181.0
999981	99.0	216.0	122.0	43.0	148.0
999982	95.0	222.0	173.0	39.0	148.0
999983	105.0	148.0	214.0	42.0	63.0

1308677 rows × 5 columns

혈당 데이터 추출

In [9]:
```
diabetes = data['식전혈당(공복혈당)']
diabetes
```

Out[9]:
```
0          99.0
1         106.0
2          98.0
3          95.0
4         101.0
          ...
999978     92.0
999979     82.0
999981     99.0
999982     95.0
999983    105.0
Name: 식전혈당(공복혈당), Length: 1308677, dtype: float64
```

인원 수 구하기

In [10]:
```
diabetes_mask = diabetes >= 126
diabetes_mask
```

Out[10]:
```
0         False
1         False
2         False
3         False
4         False
          ...
999978    False
999979    False
999981    False
999982    False
999983    False
Name: 식전혈당(공복혈당), Length: 1308677, dtype: bool
```

In [11]:
```
print("당뇨병인 사람의 수 : ",diabetes_mask.sum())
```

```
print("당뇨병이 아닌 사람의 수 : ",len(diabetes_mask)-diabetes_mask.sum())
```

```
당뇨병인 사람의 수 :  91263
당뇨병이 아닌 사람의 수 :  1217414
```

당뇨가 아닌 사람이 훨씬 많다.

In [12]:
```
Px = 91263/1308677 ; Py = 1217414/1308677
print("당뇨 확률 :",Px,"정상 확률 :",Py)
```

```
당뇨 확률 : 0.0697368410998283 정상 확률 : 0.9302631589001717
```

In [13]:
```
P = np.array([Px,Py])
P
```

Out[13]:
```
array([0.06973684, 0.93026316])
```

혈당 엔트로피

In [14]:
```
def H(p):
    id_p = np.where(p != 0)
    return -np.sum(p[id_p]*np.log2(p[id_p]))
```

In [15]:
```
parent = H(P) # 부모 엔트로피
parent
```

Out[15]:
```
0.3649408252412849
```

총콜레스테롤 엔트로피

콜레스테롤의 정상 범위는 200 미만, 높은게 240이상.
따라서 200미만, 200~239, 400이상 총 세가지 경우로 나누었다.

In [16]:
```
TC_low = data[data['총콜레스테롤'] < 200]
TC = data[(data['총콜레스테롤'] >= 200) & (data['총콜레스테롤'] < 240) ]
TC_high = data[data['총콜레스테롤']>= 240]
```

In [17]:
```
TC_low
```

Out[17]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
0	99.0	193.0	92.0	48.0	126.0
2	98.0	136.0	104.0	41.0	74.0
4	101.0	199.0	104.0	61.0	117.0
6	89.0	196.0	75.0	66.0	115.0
7	94.0	185.0	101.0	58.0	107.0
...	...	...	...	...	...
999960	82.0	178.0	125.0	48.0	105.0
999962	90.0	169.0	223.0	31.0	93.0

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
999963	115.0	157.0	110.0	64.0	71.0
999978	92.0	158.0	139.0	70.0	60.0
999983	105.0	148.0	214.0	42.0	63.0

741968 rows × 5 columns

In [18]:
```
TC
```

Out[18]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
1	106.0	228.0	121.0	55.0	148.0
3	95.0	201.0	106.0	76.0	104.0
5	99.0	218.0	232.0	77.0	95.0
8	104.0	217.0	100.0	56.0	141.0
13	82.0	200.0	77.0	55.0	129.0
...	...	...	...	...	...
999967	107.0	203.0	256.0	51.0	100.0
999974	86.0	217.0	111.0	52.0	143.0
999976	107.0	219.0	85.0	62.0	140.0
999981	99.0	216.0	122.0	43.0	148.0
999982	95.0	222.0	173.0	39.0	148.0

407574 rows × 5 columns

In [19]:
```
TC_high
```

Out[19]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
26	81.0	293.0	243.0	29.0	215.0
29	87.0	246.0	209.0	54.0	150.0
40	87.0	240.0	512.0	50.0	146.0
43	98.0	241.0	513.0	46.0	82.0
48	90.0	263.0	121.0	60.0	178.0
...	...	...	...	...	...
999933	104.0	240.0	303.0	38.0	141.0
999957	97.0	243.0	72.0	66.0	162.0
999959	108.0	266.0	125.0	60.0	181.0
999975	87.0	273.0	125.0	93.0	155.0
999979	82.0	274.0	239.0	45.0	181.0

159135 rows × 5 columns

In [20]:
```
TC_high.max()
```

Out[20]:
```
식전혈당(공복혈당)      199.0
총콜레스테롤          2386.0
트리글리세라이드        9490.0
HDL콜레스테롤         8110.0
LDL콜레스테롤         5119.0
dtype: float64
```

총 콜레스테롤이 높은 데이터를 찍어보니 2386으로 이상치가 들어있다.. 이건 나중에 처리하고 다시 구해야될 듯

In [21]:
```
print("TC_low 데이터 수 :",len(TC_low))
print("TC 데이터 수 :",len(TC))
print("TC_high 데이터 수 :",len(TC_high))
print(len(TC_low)+len(TC)+len(TC_high))
```

```
TC_low 데이터 수 : 741968
TC 데이터 수 : 407574
TC_high 데이터 수 : 159135
1308677
```

확률 구하기

In [22]:
```
TC_low = TC_low['식전혈당(공복혈당)']
TC = TC['식전혈당(공복혈당)']
TC_high = TC_high['식전혈당(공복혈당)']
```

In [23]:
```
TC_low_diabetes_mask = TC_low >= 126
TC_diabetes_mask = TC >= 126
TC_high_diabetes_mask = TC_high >= 126
```

In [24]:
```
TC_low_diabetes_mask
```

Out[24]:
```
0         False
2         False
4         False
6         False
7         False
          ...
999960    False
999962    False
999963    False
999978    False
999983    False
Name: 식전혈당(공복혈당), Length: 741968, dtype: bool
```

In [26]:
```
print('TC_low 당뇨병 사람 수 :',TC_low_diabetes_mask.sum() , "  정상 수:", len(TC_
print('TC 당뇨병 사람 수 :',TC_diabetes_mask.sum(), "  정상 수:",len(TC_diabetes_m
print('TC_high 당뇨병 사람 수 :',TC_high_diabetes_mask.sum(), "  정상 수:",len(TC_
```

```
TC_low 당뇨병 사람 수 : 58504   정상 수: 683464
TC 당뇨병 사람 수 : 21909   정상 수: 385665
TC_high 당뇨병 사람 수 : 10850   정상 수: 148285
```

In [27]:
```
Px = [58504/741968, 21909/407574, 10850/159135] ; Py = [683464/741968, 385665
print("Px :",Px, "\nPy :",Py)
```

```
Px : [0.07884976171479094, 0.053754655596284354, 0.06818110409400824]
Py : [0.921150238285209, 0.9462453444037157, 0.9318188959059918]
```

엔트로피 구하기

In [28]:
```
TC_entropy = []
for x,y in zip(Px,Py):
    P = np.array([x,y])
    print(P)
    print(H(P))
    TC_entropy.append(H(P))
TC_entropy
```

```
[0.07884976 0.92115024]
0.39811323009720984
[0.05375466 0.94624534]
0.3021372683238905
[0.0681811 0.9318189]
0.3590989316825833
```

Out[28]: [0.39811323009720984, 0.3021372683238905, 0.3590989316825833]

In [29]:
```
print("TC_low 엔트로피 :",TC_entropy[0])
print("TC 엔트로피 :",TC_entropy[1])
print("TC_high 엔트로피 :",TC_entropy[2])
```

```
TC_low 엔트로피 : 0.39811323009720984
TC 엔트로피 : 0.3021372683238905
TC_high 엔트로피 : 0.3590989316825833
```

정보증가량 구하기

부모 엔트로피 - 각각 자식(자식 엔트로피가 될 확률*자식 엔트로피)의 합

부모 엔트로피 : parent

TC_low 확률 : 741968/1308677

TC 확률 : 407574/1308677

TC_high 확률 : 159135/1308677

In [30]:
```
TC_Px = np.array([741968/1308677 , 407574/1308677,  159135/1308677])
TC_Px
```

Out[30]: array([0.56696037, 0.31143972, 0.12159991])

In [31]:
```
TC_IG = parent - sum(TC_Px * TC_entropy)
print("총콜레스테롤 IG :",TC_IG)
```

```
총콜레스테롤 IG : 0.001462456933508549
```

정보 증가량이 매우 낮게 나왔다.

트리글리세라이드 엔트로피

150 미만이면 정상, 150~199면 주의, 200 이상이면 치료가 필요한 수준

In [32]:
```
TG_low = data[data['트리글리세라이드'] < 150]
TG = data[(data['트리글리세라이드'] >= 150) & (data['트리글리세라이드'] < 200) ]
TG_high = data[data['트리글리세라이드']>= 200]
```

In [33]:
```
TG_low
```

Out[33]:

	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드	HDL콜레스테롤	LDL콜레스테롤
0	99.0	193.0	92.0	48.0	126.0
1	106.0	228.0	121.0	55.0	148.0
2	98.0	136.0	104.0	41.0	74.0
3	95.0	201.0	106.0	76.0	104.0
4	101.0	199.0	104.0	61.0	117.0
...	...	...	...	...	...
999974	86.0	217.0	111.0	52.0	143.0
999975	87.0	273.0	125.0	93.0	155.0
999976	107.0	219.0	85.0	62.0	140.0
999978	92.0	158.0	139.0	70.0	60.0
999981	99.0	216.0	122.0	43.0	148.0

942937 rows × 5 columns

In [34]:
```
TG_high.max()
```

Out[34]:
```
식전혈당(공복혈당)      199.0
총콜레스테롤          2386.0
트리글리세라이드        9490.0
HDL콜레스테롤         8110.0
LDL콜레스테롤         5119.0
dtype: float64
```

역시나 제일 높은 수가 비정상적으로 나왔다.

In [36]:
```
TG_low = TG_low['식전혈당(공복혈당)']
TG = TG['식전혈당(공복혈당)']
TG_high = TG_high['식전혈당(공복혈당)']
```

In [37]:
```
print("TG_low 데이터 수 :",len(TG_low))
print("TG 데이터 수 :",len(TG))
print("TG_high 데이터 수 :",len(TG_high))
print(len(TG_low)+len(TG)+len(TG_high))
```

```
TG_low 데이터 수 : 942937
TG 데이터 수 : 178141
TG_high 데이터 수 : 187599
1308677
```

In [38]:
```
TG_low_diabetes_mask = TG_low >= 126
TG_diabetes_mask = TG >= 126
TG_high_diabetes_mask = TG_high >= 126
```

In [39]:
```
print('TG_low 당뇨병 사람 수 :',TG_low_diabetes_mask.sum() , "  정상 수:", len(TG_
print('TG 당뇨병 사람 수 :',TG_diabetes_mask.sum(), "  정상 수:",len(TG_diabetes_m
print('TG_high 당뇨병 사람 수 :',TG_high_diabetes_mask.sum(), "  정상 수:",len(TG_
```

```
TG_low 당뇨병 사람 수 : 50727   정상 수: 892210
TG 당뇨병 사람 수 : 17054   정상 수: 161087
TG_high 당뇨병 사람 수 : 23482   정상 수: 164117
```

In [40]:
```
Px = [50727/942937, 17054/178141, 23482/187599] ; Py = [892210/942937, 161087
print("Px :",Px, "\nPy :",Py)
```

```
Px : [0.05379680720981359, 0.09573315519728755, 0.12517124291707313]
Py : [0.9462031927901864, 0.9042668448027125, 0.8748287570829268]
```

엔트로피 구하기

In [41]:
```
TG_entropy = []
for x,y in zip(Px,Py):
    P = np.array([x,y])
    print(P)
    print(H(P))
    TG_entropy.append(H(P))
TG_entropy
```

```
[0.05379681 0.94620319]
0.30231165608480187
[0.09573316 0.90426684]
0.45532220951561375
[0.12517124 0.87482876]
0.5440449895234732
```

Out[41]: `[0.30231165608480187, 0.45532220951561375, 0.5440449895234732]`

In [42]:
```
print("TG_low 엔트로피 :",TG_entropy[0])
print("TG 엔트로피 :",TG_entropy[1])
print("TG_high 엔트로피 :",TG_entropy[2])
```

```
TG_low 엔트로피 : 0.30231165608480187
TG 엔트로피 : 0.45532220951561375
TG_high 엔트로피 : 0.5440449895234732
```

정보증가량 구하기

부모 엔트로피 : parent TG_low 확률 : 942937/1308677

TG 확률 : 178141/1308677

TG_high 확률 : 187599/1308677

In [44]:
```
TG_Px = np.array([942937/1308677, 178141/1308677,  187599/1308677])
TG_Px
```

Out[44]: `array([0.72052691, 0.13612297, 0.14335012])`

In [45]:
```
TG_IG = parent - sum(TG_Px * TG_entropy)
print("트리글리세라이드 IG :",TG_IG)
```

```
트리글리세라이드 IG : 0.007148416748914532
```

엄청 낮게 나왔다..

HDL콜레스테롤

60이상이 양호한 것이고 40 이하는 나쁨수준이다.

In [47]:
```
HC_low = data[data['HDL콜레스테롤'] < 40]
HC = data[(data['HDL콜레스테롤'] >= 40) & (data['HDL콜레스테롤'] < 60) ]
HC_high = data[data['HDL콜레스테롤']>= 60]
```

In [48]:
```
HC_high.max()
```

Out[48]:
```
식전혈당(공복혈당)      199.0
총콜레스테롤          2033.0
트리글리세라이드        9490.0
HDL콜레스테롤         8110.0
LDL콜레스테롤         5119.0
dtype: float64
```

역시나 최대값에 문제가 있다.

In [49]:
```
HC_low = HC_low['식전혈당(공복혈당)']
HC = HC['식전혈당(공복혈당)']
HC_high = HC_high['식전혈당(공복혈당)']
```

In [50]:
```
print("HC_low 데이터 수 :",len(HC_low))
print("HC 데이터 수 :",len(HC))
print("HC_high 데이터 수 :",len(HC_high))
print(len(HC_low)+len(HC)+len(HC_high))
```

```
HC_low 데이터 수 : 125578
HC 데이터 수 : 683328
HC_high 데이터 수 : 499771
1308677
```

In [51]:
```
HC_low_diabetes_mask = HC_low >= 126
HC_diabetes_mask = HC >= 126
HC_high_diabetes_mask = HC_high >= 126
```

In [52]:
```
HC_low_diabetes_mask
```

Out[52]:
```
11          True
22         False
25         False
26         False
76          True
           ...
999888     False
999897     False
999933     False
999962     False
999982     False
Name: 식전혈당(공복혈당), Length: 125578, dtype: bool
```

In [53]:
```
print('HC_low 당뇨병 사람 수 :',HC_low_diabetes_mask.sum() , "  정상 수:", len(HC_
print('HC 당뇨병 사람 수 :',HC_diabetes_mask.sum(), "  정상 수:",len(HC_diabetes_ma
print('HC_high 당뇨병 사람 수 :',HC_high_diabetes_mask.sum(), "  정상 수:",len(HC_
```

```
HC_low 당뇨병 사람 수 : 15066   정상 수: 110512
HC 당뇨병 사람 수 : 53928   정상 수: 629400
HC_high 당뇨병 사람 수 : 22269   정상 수: 477502
```

In [54]:
```
Px = [15066/125578, 53928/683328, 22269/499771] ; Py = [110512/125578, 629400
print("Px :",Px, "\nPy :",Py)
```

```
Px : [0.11997324372103393, 0.07891964034841248, 0.044558407750749845]
Py : [0.8800267562789661, 0.9210803596515875, 0.9554415922492502]
```

엔트로피 구하기

In [55]:
```
HC_entropy = []
for x,y in zip(Px,Py):
    P = np.array([x,y])
    print(P)
    print(H(P))
    HC_entropy.append(H(P))
HC_entropy
```

```
[0.11997324 0.88002676]
0.529283950299191
[0.07891964 0.92108036]
0.39836098929054375
[0.04455841 0.95544159]
0.2628154309621035
```

Out[55]:
```
[0.529283950299191, 0.39836098929054375, 0.2628154309621035]
```

In [56]:
```
print("HC_low 엔트로피 :",HC_entropy[0])
print("HC 엔트로피 :",HC_entropy[1])
print("HC_high 엔트로피 :",HC_entropy[2])
```

```
HC_low 엔트로피 : 0.529283950299191
HC 엔트로피 : 0.39836098929054375
HC_high 엔트로피 : 0.2628154309621035
```

정보 증가량 구하기

부모 엔트로피 : parent

HC_low 확률 : 125578/1308677

HC 확률 : 683328/1308677

HC_high 확률 : 499771/1308677

In [57]:
```
HC_Px = np.array([125578/1308677, 683328/1308677,  499771/1308677])
HC_Px
```

Out[57]:
```
array([0.09595798, 0.52215176, 0.38189026])
```

In [58]:
```
HC_IG = parent - sum(HC_Px * HC_entropy)
print("HDL콜레스테롤 IG :",HC_IG)
```

```
HDL콜레스테롤 IG : 0.005780261750093496
```

잘못 구하고 있는건가..?

LDL 콜레스테롤

당뇨병이 있으면 100미만, 아니면 130미만이 정상이고 160이상이면 높은것

In [59]:
```
LC_low = data[data['LDL콜레스테롤'] < 130]
```

```
LC = data[(data['LDL콜레스테롤'] >= 130) & (data['LDL콜레스테롤'] < 160) ]
LC_high = data[data['LDL콜레스테롤']>= 160]
```

In [60]:
```
LC_high.max()
```

Out[60]:
```
식전혈당(공복혈당)      199.0
총콜레스테롤          2386.0
트리글리세라이드        9490.0
HDL콜레스테롤         8110.0
LDL콜레스테롤         5119.0
dtype: float64
```

역시나 비정상이다.

In [62]:
```
LC_low = LC_low['식전혈당(공복혈당)']
LC = LC['식전혈당(공복혈당)']
LC_high = LC_high['식전혈당(공복혈당)']
```

In [63]:
```
print("LC_low 데이터 수 :",len(LC_low))
print("LC 데이터 수 :",len(LC))
print("LC_high 데이터 수 :",len(LC_high))
print(len(LC_low)+len(LC)+len(LC_high))
```

```
LC_low 데이터 수 : 915688
LC 데이터 수 : 271543
LC_high 데이터 수 : 121446
1308677
```

In [64]:
```
LC_low_diabetes_mask = LC_low >= 126
LC_diabetes_mask = LC >= 126
LC_high_diabetes_mask = LC_high >= 126
```

In [65]:
```
LC_low_diabetes_mask
```

Out[65]:
```
0         False
2         False
3         False
4         False
5         False
          ...
999962    False
999963    False
999967    False
999978    False
999983    False
Name: 식전혈당(공복혈당), Length: 915688, dtype: bool
```

In [66]:
```
print('LC_low 당뇨병 사람 수 :',LC_low_diabetes_mask.sum() , "  정상 수:", len(LC_
print('LC 당뇨병 사람 수 :',LC_diabetes_mask.sum(), "  정상 수:",len(LC_diabetes_m
print('LC_high 당뇨병 사람 수 :',LC_high_diabetes_mask.sum(), "  정상 수:",len(LC_
```

```
LC_low 당뇨병 사람 수 : 69910   정상 수: 845778
LC 당뇨병 사람 수 : 14040   정상 수: 257503
LC_high 당뇨병 사람 수 : 7313   정상 수: 114133
```

In [67]:
```
Px = [69910/915688, 14040/271543, 7313/121446] ; Py = [845778/915688, 257503/
print("Px :",Px, "\nPy :",Py)
```

```
Px : [0.07634696534190685, 0.051704518253094356, 0.06021606310623652]
Py : [0.9236530346580931, 0.9482954817469057, 0.9397839368937635]
```

엔트로피 구하기

In [69]:
```
LC_entropy = []
for x,y in zip(Px,Py):
    P = np.array([x,y])
    print(P)
    print(H(P))
    LC_entropy.append(H(P))
LC_entropy
```

```
[0.07634697 0.92365303]
0.3891748456581644
[0.05170452 0.94829548]
0.293593972227775
[0.06021606 0.93978394]
0.3283020125522487
```

Out[69]: [0.3891748456581644, 0.293593972227775, 0.3283020125522487]

In [70]:
```
print("LC_low 엔트로피 :",HC_entropy[0])
print("LC 엔트로피 :",HC_entropy[1])
print("LC_high 엔트로피 :",HC_entropy[2])
```

```
LC_low 엔트로피 : 0.529283950299191
LC 엔트로피 : 0.39836098929054375
LC_high 엔트로피 : 0.2628154309621035
```

정보증가량 구하기

부모 엔트로피 : parent

LC_low 확률 : 915688/1308677

LC 확률 : 271543/1308677

LC_high 확률 : 121446/1308677

In [72]:
```
LC_Px = np.array([915688/1308677, 271543/1308677,  121446/1308677])
LC_Px
```

Out[72]: array([0.69970512, 0.20749429, 0.09280059])

In [73]:
```
LC_IG = parent - sum(LC_Px * LC_entropy)
print("LDL콜레스테롤 IG :",LC_IG)
```

```
LDL콜레스테롤 IG : 0.0012474996245741
```

찾아봤을 때는 LDL콜레스테롤과 중성지방이 당뇨병과 연관이 크다고 했었는데 IG는 엄청 낮다...

제대로 구한건지 모르겠네

너무 어렵다

In [ ]:
```

```