In [3]:
```
import numpy as np
import pandas as pd
import math
data2017 = pd.read_csv('NHIS_OPEN_GJ_2017_100.csv', encoding='euc-kr')
data2018 = pd.read_csv('NHIS_OPEN_GJ_2018_100.csv')
```

In [276…
```
data2017.head()
```

Out[276…

	기준년도	가입자일련번호	성별코드	연령대코드(5세단위)	시도코드	신장(5Cm단위)	체중(5Kg단위)	허리둘레	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압	이완기혈압	식전혈당(공복혈당)	총콜레스테롤	트리서0
0	2017	1	1	8	43	170	75	90.0	1.0	1.0	1.0	1.0	120.0	80.0	99.0	193.0	92
1	2017	2	1	7	11	180	80	89.0	0.9	1.2	1.0	1.0	130.0	82.0	106.0	228.0	12
2	2017	3	1	9	41	165	75	91.0	1.2	1.5	1.0	1.0	120.0	70.0	98.0	136.0	104
3	2017	4	1	11	48	175	80	91.0	1.5	1.2	1.0	1.0	145.0	87.0	95.0	201.0	106
4	2017	5	1	11	30	165	60	80.0	1.0	1.2	1.0	1.0	138.0	82.0	101.0	199.0	104

In [277…
```
data2018.head()
```

Out[277…

	기준년도	가입자일련번호	성별코드	연령대코드(5세단위)	시도코드	신장(5Cm단위)	체중(5Kg단위)	허리둘레	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압	이완기혈압	식전혈당(공복혈당)	총콜레스테롤	트글세이
0	2018	1	2	7	48	160	60	79.5	1.5	1.5	1.0	1.0	110.0	60.0	96.0	NaN	Na
1	2018	2	1	6	26	170	55	69.3	1.2	0.8	1.0	1.0	128.0	78.0	79.0	NaN	Na
2	2018	3	1	12	28	165	70	85.0	0.8	0.8	2.0	1.0	128.0	65.0	80.0	NaN	Na
3	2018	4	2	15	27	150	45	71.5	0.4	0.3	1.0	1.0	151.0	89.0	100.0	234.0	90.
4	2018	5	2	14	41	145	50	77.0	0.7	0.6	1.0	1.0	114.0	62.0	124.0	NaN	Na

In [4]:
```
data = pd.concat([data2017, data2018])
```

In [279…
```
len(data)
```

Out[279… 2000000

In [280…
```
data.tail()
```

Out[280…

	기준년도	가입자일련번호	성별코드	연령대코드(5세단위)	시도코드	신장(5Cm단위)	체중(5Kg단위)	허리둘레	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압	이완기혈압	식전혈당(공복혈당
999995	2018	999996	2	11	41	165	75	84.0	1.2	1.2	1.0	1.0	110.0	70.0	107.(
999996	2018	999997	1	7	41	170	75	88.0	2.0	2.0	1.0	1.0	136.0	88.0	114.(
999997	2018	999998	1	8	41	175	80	87.0	1.2	1.2	1.0	1.0	162.0	90.0	98.(
999998	2018	999999	1	11	41	165	70	80.2	0.9	1.2	1.0	1.0	140.0	98.0	94.(
999999	2018	1000000	1	5	47	165	60	70.0	1.5	1.2	1.0	1.0	120.0	80.0	85.(

In [8]:

```
data = data[0 <= data['식전혈당(공복혈당)']]
data =  data[300 >= data['식전혈당(공복혈당)']]
data = data.reset_index(drop=True, inplace=False)
data.tail()
```

Out[8]:

	기준년도	가입자일련번호	성별코드	연령대코드(5세단위)	시도코드	신장(5Cm단위)	체중(5Kg단위)	허리둘레	시력(좌)	시력(우)	…	감마지티피	흡연상태	음주여부	구강검진수검여부	치우증
1990974	2018	999996	2	11	41	165	75	84.0	1.2	1.2	…	19.0	1.0	NaN	0	Na
1990975	2018	999997	1	7	41	170	75	88.0	2.0	2.0	…	90.0	1.0	1.0	1	0
1990976	2018	999998	1	8	41	175	80	87.0	1.2	1.2	…	36.0	2.0	1.0	0	Na
1990977	2018	999999	1	11	41	165	70	80.2	0.9	1.2	…	14.0	2.0	NaN	0	Na
1990978	2018	1000000	1	5	47	165	60	70.0	1.5	1.2	…	11.0	1.0	1.0	1	0

5 rows × 34 columns

In [282…

```
add_col = data['식전혈당(공복혈당)'].copy()
data_U = data['요단백'].copy()
data_V = data['혈청크레아티닌'].copy()
data_ZZ = data['음주여부'].copy()
data_ZZ[data_ZZ.isnull()] = 0
```

In [283…

```
add_col[add_col < 126] = 0
add_col[add_col >= 126] = 1

data['당뇨병'] = add_col
```

In [284…

```
def getEntropy(n, all):
    Px1 = n/all
    Px2 = (all-n)/all
```

```
    Hx = (-Px1 * np.log2(Px1)) + (-Px2 * np.log2(Px2))
    return Hx
```

In [285…

```
DnCnt = data[data['당뇨병']==1]['당뇨병'].sum()
All = len(data)
Hx = getEntropy(DnCnt, All)
print(Hx)
```

```
0.390673901150853
```

In [286…

```
data.tail()
```

Out[286…

	기준년도	가입자일련번호	성별코드	연령대코드(5세단위)	시도코드	신장(5Cm단위)	체중(5Kg단위)	허리둘레	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압	이완기혈압	식혈(공
1990974	2018	999996	2	11	41	165	75	84.0	1.2	1.2	1.0	1.0	110.0	70.0	107
1990975	2018	999997	1	7	41	170	75	88.0	2.0	2.0	1.0	1.0	136.0	88.0	114
1990976	2018	999998	1	8	41	175	80	87.0	1.2	1.2	1.0	1.0	162.0	90.0	98
1990977	2018	999999	1	11	41	165	70	80.2	0.9	1.2	1.0	1.0	140.0	98.0	94
1990978	2018	1000000	1	5	47	165	60	70.0	1.5	1.2	1.0	1.0	120.0	80.0	85

In [287…

```
def digitSegment(parent, unit):
    child = {}
    unit = unit*10
    min_value = parent.min() * 10
    max_value = parent.max() * 10
    start = int(min_value//unit)
    end = int(max_value//unit)
    for i in range(start, end+1):
        key = str(i)
        child[key] = [[],[]]
    for i in range(All):
        if math.isnan(parent[i]): continue
        dt = parent[i] * 10
        key = int(dt//unit)
        key = str(key)
        if(data['당뇨병'][i] == 0): child[key][0].append(parent[i])
        else: child[key][1].append(parent[i])
    return child
```

In [288…

```
def categorySegment(parent):
    child = {}
    min_value = int(parent.min())
    max_value = int(parent.max())
    for i in range(min_value, max_value+1):
        key = str(i)
        child[key] = [[],[]]
    for i in range(All):
        if math.isnan(parent[i]): continue
        key = int(parent[i])
```

```
        key = str(key)
        if(data['당뇨병'][i] == 0): child[key][0].append(parent[i])
        else: child[key][1].append(parent[i])

    return child
```

In [289…

```
def ProbEnt(segment):
    SegList = []
    total = 0
    totalDN = 0
    totalNDN = 0
    for key, value in segment.items():
        NDNcnt = len(value[0])
        DNcnt = len(value[1])
        totalDN += DNcnt
        totalNDN += NDNcnt
        subtotal = NDNcnt + DNcnt
        if(subtotal == 0): continue
        total += subtotal
        if(DNcnt == 0 or NDNcnt == 0):
            SegList.append([subtotal, 1])
            continue
        entropy = getEntropy(DNcnt, subtotal)
        SegList.append([subtotal, entropy])

    for i in range(len(SegList)):
        SegList[i][0] = SegList[i][0] / total

    SegList.append([total, totalDN])
    return SegList
```

In [290…

```
def getIG(entropy, seglist):
    answer = entropy
    for seg in seglist[:-1]:
        m = seg[0]*seg[1]
        answer -= m
    return answer
```

In [291…

```
dt = categorySegment(data_U)
seglist = ProbEnt(dt)
Hx = getEntropy(seglist[-1][1], seglist[-1][0])
print("요단백 엔트로피:", Hx)
print("--- 분화 후 ---")
print("요단백 정보증가량:", getIG(Hx, seglist))
```

```
요단백 엔트로피: 0.39017200343397407
--- 분화 후 ---
요단백 정보증가량: 0.0052101303447348035
```

In [292…

```
dt = digitSegment(data_V, 0.2)
seglist = ProbEnt(dt)
Hx = getEntropy(seglist[-1][1], seglist[-1][0])
print("혈청크레아티닌 엔트로피:", Hx)
print("--- 분화 후 ---")
print("혈청크레아티닌 정보증가량:", getIG(Hx, seglist))
```

```
혈청크레아티닌 엔트로피: 0.39067445453757704
--- 분화 후 ---
혈청크레아티닌 정보증가량: 0.0027044415530841992
```

In [293…

```
dt = categorySegment(data_ZZ)
seglist = ProbEnt(dt)
Hx = getEntropy(seglist[-1][1], seglist[-1][0])
print("음주여부 엔트로피:", Hx)
print("--- 분화 후 ---")
print("음주여부 정보증가량:", getIG(Hx, seglist))
```

```
음주여부 엔트로피: 0.390673901150853
--- 분화 후 ---
음주여부 정보증가량: 0.00020317736750488646
```