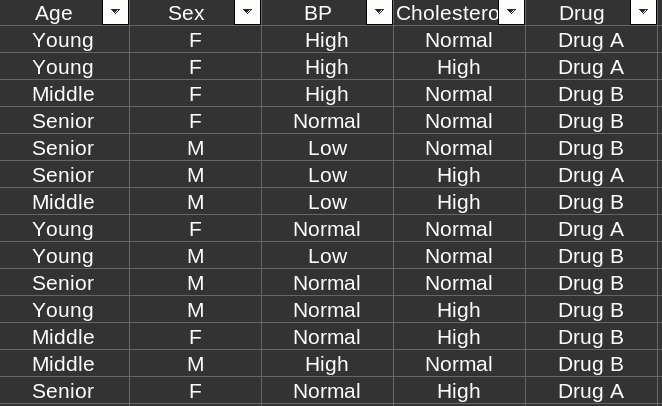
## ENTROPY

## 2. Hãy xây dựng và vẽ cây quyết định (Từng bước) bằng thuật toán xây dựng cây quyết định bằng hàm Entropy ?

### - (1) tìm Root



**B1: tính entropy tổng của labels (kết quả)**

Entropy(S-label) = -[Tần suất \* log2(DrugA) + Tần suất \* log2(DrugB)]

= -[(5/14)\*log2(5/14)+(9/14)\*log2(9/14)]

= 0.94

**B2: tính average entropy**

Entroy(Young) = A(3) + B(2)

= -[(3/5)\*log2(3/5)+(2/5)\*log2(2/5)] = 0.971

Entroy(Middle) = A(0) + B(4)

= -[(0/4)\*log2(0/4)+(4/4)\*log2(4/4)] = 0

Entroy(Senior) = A(2) + B(3)

= -[(2/5)\*log2(2/5)+(3/5)\*log2(3/5)] = 0.971

**AE(Age) = 5/14\*Entroy(Young)+4/14\*Entroy(Middle)+5/14\*Entroy(Senior)**

**= 5/14\*0.971+4/14\*0+5/14\*0.971**

**= 0.694**

Entroy(F) = A(4) + B(3)

= -[(4/7)\*log2(4/7)+(3/7)\*log2(3/7)] = 0.985

Entroy(M) = A(1) + B(6)

= -[(1/7)\*log2(1/7)+(6/7)\*log2(6/7)] = 0.592

**AE(Sex) = 7/14\*Entroy(F)+7/14\*Entroy(M)**

**= 7/14\*0.985+7/14\*0.592**

**= 0.789**

Entroy(Low) = A(1) + B(3)

= -[(1/4)\*log2(1/4)+(3/4)\*log2(3/4)] = 0.811

Entroy(Normal) = A(2) + B(4)

= -[(2/6)\*log2(2/6)+(4/6)\*log2(4/6)] = 0.918

Entroy(High) = A(2) + B(2)

= -[(2/4)\*log2(2/4)+(2/4)\*log2(2/4)] = 1

**AE(BP) = 4/14\*Entroy(Low)+6/14\*Entroy(Normal)+4/14\*Entroy(High)**

**= 4/14\*0.811+6/14\*0.918+4/14\*1**

**= 0.911**

Entroy(Chol-A) = Normal(2) + High(3)

= -[(2/5)\*log2(2/5)+(3/5)\*log2(3/5)]

= 0.971

Entroy(Chol-B) = Normal(6) + High(3)

= -[(6/9)\*log2(6/9)+(3/9)\*log2(3/9)]

= 0.918

**AE(Chol) = 5/14\*Entroy(Chol-A) + 9/14\*Entroy(Chol-B)**

**= 5/14(0.971)+9/14(0.918)**

**= 0.937**

**B3: Tính tới chỉ số thông tin mang lại của từng thuộc tính**

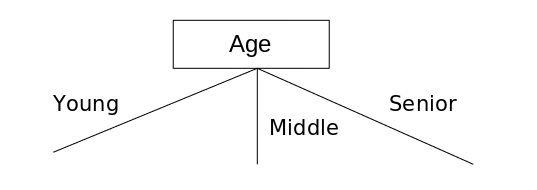
Gain(A) = Entropy(S-label) – AE(S)

**Gain(Age) = 0.94 - 0.694 = 0.246 chọn root**

Gain(Sex) = 0.94 - 0.789 = 0.151

Gain(BP) = 0.94 - 0.911 = 0.029

Gain(Chol) = 0.94 - 0.892 = 0.048

****

### - (1-1) Tìm node nhánh Young

****

**B1: tính entropy tổng của labels (kết quả)**

Entropy(S-label) = A(3) + B(2)

= -[(3/5)\*log2(3/5)+(2/5)\*log2(2/5)]

**= 0.971**

**B2: tính average entropy**

Entroy(F) = A(3) + B(0)

= -[(3/3)\*log2(3/3)+(0/3)\*log2(0/3)] = 0

Entroy(M) = A(0) + B(2)

= -[(0/2)\*log2(0/2)+(2/2)\*log2(2/2)] = 0

**AE(Sex) = 3/5\*Entroy(F)+2/5\*Entroy(M)**

**= 3/5\*0+2/5\*0 = 0**

Entroy(Low) = A(1) + B(0)

= -[(1/1)\*log2(1/1)+(0/1)\*log2(0/1)] = 0

Entroy(Normal) = A(1) + B(1)

= -[(1/2)\*log2(1/2)+(1/2)\*log2(1/2)] = 1

Entroy(High) = A(2) + B(0)

= -[(2/2)\*log2(2/2)+(0/2)\*log2(0/2)] = 0

**AE(BP) = 1/5\*Entroy(Low)+2/5\*Entroy(Normal)+2/5\*Entroy(High)**

**= 1/5\*0+2/5\*1+2/5\*0 = 0.4**

Entroy(Normal) = A(2) + B(1)

= -[(2/3)\*log2(2/3)+(1/3)\*log2(1/3)] = 0.918

Entroy(High) = A(1) + B(1)

= -[(1/2)\*log2(1/2)+(1/2)\*log2(1/2)] = 1

**AE(Chol) = 3/5\*Entroy(Normal)+5/\*Entroy(High)**

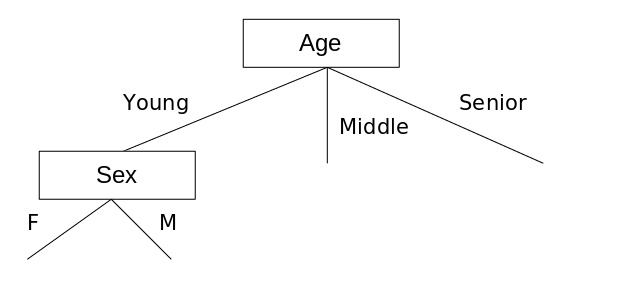
**= 3/5\*0.918+2/5\*1 = 0.951**

**B3: Tính tới chỉ số thông tin mang lại của từng thuộc tính**

**Gain(Sex) = 0.971 - 0 = 0.971 Chon Node**

Gain(BP) = 0.971 - 0.4 = 0.571

Gain(Chol) = 0.971 - 0.951 = 0.02



### - (1-1-1) Tìm node nhánh Young-F

****

**B1: tính entropy tổng của labels (kết quả)**

Entroy(S-label) = A(3) + B(0)

= -[(3/3)\*log2(3/3)+(0/3)\*log2(0/3)]

= 0

=> Không phải node, mà là lá, với label Drug A

### 

### - (1-1-2) Tìm node nhánh Young-M

****

Entroy(S-label) = A(0) + B(2)

= -[(0/2)\*log2(0/2)+(2/2)\*log2(2/2)]

= 0

**=> Không phải node, mà là lá, với label Drug B**

### 

### - (1-2) Tìm node nhánh Middle

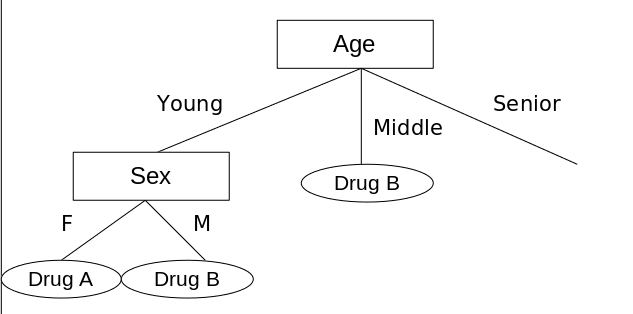
****

Entroy(S-label) = A(0) + B(4)

= -[(0/4)\*log2(0/4)+(4/4)\*log2(4/4)]

= 0

**=> Không phải node, mà là lá, với label Drug B**

****

### - (1-3) Tìm node nhánh Senior

****

**B1: tính entropy tổng của labels (kết quả)**

Entroy(S-label) = A(2) + B(5)

= -[(2/5)\*log2(2/5)+(3/5)\*log2(3/5)]

= 0.971

**B2: tính average entropy**

Entroy(F) = A(1) + B(1)

= -[(1/2)\*log2(1/2)+(1/2)\*log2(1/2)] = 1

Entroy(M) = A(1) + B(2)

= -[(1/3)\*log2(1/3)+(2/3)\*log2(2/3)] = 0.918

**AE(Sex) = 2/5\*Entroy(F)+3/5\*Entroy(M)**

**= 2/5\*1+3/5\*0.918**

**= 0.951**

Entroy(Low) = A(1) + B(1)

= -[(1/2)\*log2(1/2)+(1/2)\*log2(1/2)] = 1

Entroy(Normal) = A(1) + B(2)

= -[(1/3)\*log2(1/3)+(2/3)\*log2(2/3)] = 0.918

Entroy(High) = A(0) + B(0)

= -[(0/0)\*log2(0/0)+(0/0)\*log2(0/0)] = 0

**AE(BP) = 2/5\*Entroy(Low)+3/5\*Entroy(Normal)+0/5\*Entroy(High)**

**= 2/5\*1+3/5\*0.918+0/5\*0**

**= 0.951**

Entroy(Normal) = A(0) + B(0)

= -[(0/0)\*log2(0/0)+(0/0)\*log2(0/0)] = 0

Entroy(High) = A(2) + B(0)

= -[(2/2)\*log2(2/2)+(0/2)\*log2(0/2)] = 0

**AE(Chol) = 3/5\*Entroy(Normal)+5/\*Entroy(High)**

**= 3/5\*0+2/5\*0**

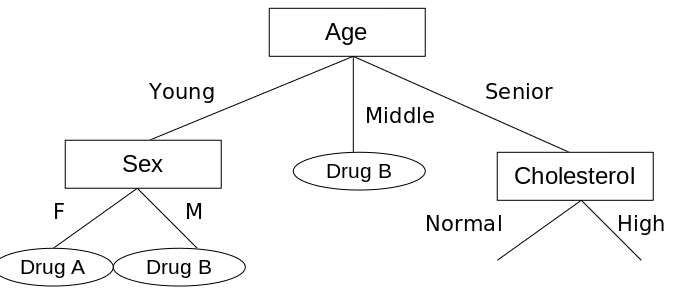
**= 0**

**B3: Tính tới chỉ số thông tin mang lại của từng thuộc tính**

Gain(Sex) = 0.971 - 0.951 = 0.02

Gain(BP) = 0.971 - 0.951 = 0.02

**Gain(Chol) = 0.971 - 0 = 0.971 Chon Node**



### - (1-3-1) Tìm node nhánh Senior-Normal

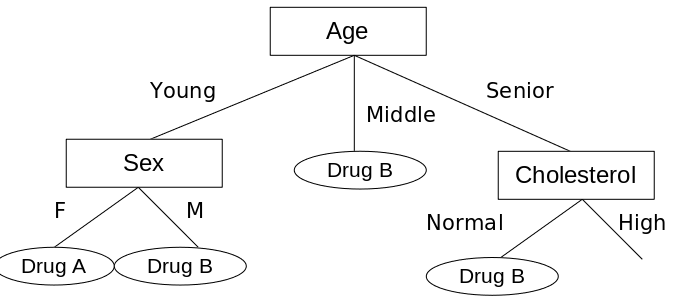
****

Entroy(S-label) = A(0) + B(3)

= -[(0/3)\*log2(0/3)+(3/3)\*log2(3/3)]

= 0

**=> Không phải node, mà là lá, với label Drug B**



### - (1-3-1) Tìm node nhánh Senior-Normal

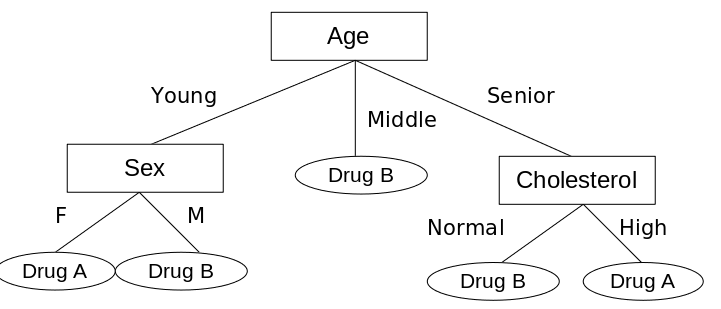
****

Entroy(S-label) = A(0) + B(3)

= -[(2/2)\*log2(2/2)+(0/2)\*log2(0/2)]

= 0

**=> Không phải node, mà là lá, với label Drug A**



## 5. Sử dụng cây quyết định Entropy để dự đoán lại dataset?

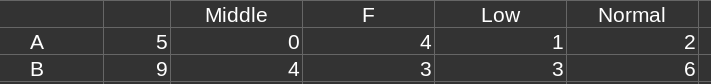
## 7. Bao nhiêu phần trăm kết quả của câu 5 giống vs dataset ban đầu? vì sao

100% => vì công thức áp dụng trên toàn tệp (chưa có tách train và test)

# XÁC SUẤT CÓ ĐIỀU KIỆN

## 3. Hãy phân loại bệnh nhân thứ 15, và dự đoán thuốc của bệnh nhân này bằng Thuật Toán Native Bayes?

### - Thống kê thuộc tính



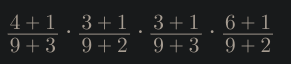
### - Tính xác suất xảy ra A hoặc B?

P(A|X) = [P(X|A) \* P(A)] / P(X)

P(B|X) = [P(X|B) \* P(B)] / P(X)

- Tính P(X|A) và P(X|B)

P(X|A) =

 =

P(X|B) =

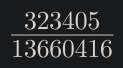
=

- Tính P(X)

P(X) = [P(X|A) \* P(A)] + [P(X|B) \* P(B)]



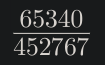
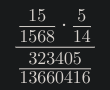
=



=

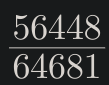
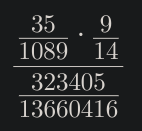
- Tính P(A|X) và P(B|X)

P(A|X) = [P(X|A) \* P(A)] / P(X)

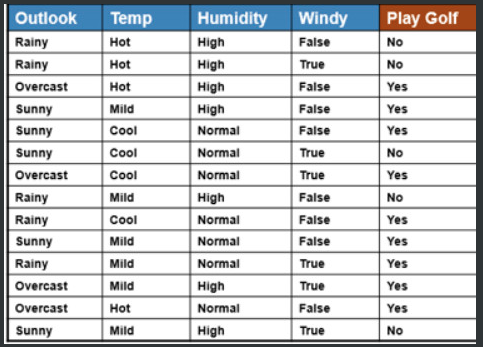


= =

P(B|X) = [P(X|B) \* P(B)] / P(X)



= =





P(X|No) = 3/5\*2/5\*1/5\*3/5 = 18/625

P(X|Yes) = 2/9\*2/9\*6/9\*3/9 = 8/729

P(X) = [P(X|No) \* P(No)] + [P(X|Yes) \* P(Yes)]

= 18/625 \* 5/14 + 8/729 \* 9/14

= 1229/70875

P(No|X) = [P(X|No) \* P(No)] / P(X)

= 18/625 \* 5/14 / (1229/70875)

= 729/1229

P(Yes|X) = [P(X|Yes) \* P(Yes)] / P(X)

= 8/729\*9/14/(1229/70875)

= 500/1229

=> Nooooooo