**Note: do tránh mất thời gian gõ nên log2 (a) sẽ được viết thành log(a).**

**Bài 1: Cho dữ liệu thu thập được như sau:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tid** | **Refund** | **Marital Status** | **Taxable income** | **Cheat** |
| 1 | No | Single | 75K | No |
| 2 | Yes | Married | 50K | No |
| 3 | No | Single | 75K | No |
| 4 | No | Married | 150K | Yes |
| 5 | No | Single | 40K | No |
| 6 | No | Married | 80K | Yes |
| 7 | No | Single | 75K | No |
| 8 | Yes | Married | 50K | No |
| 9 | Yes | Married | 50K | No |
| 10 | No | Married | 150K | Yes |
| 11 | No | Single | 40K | No |
| 12 | No | Married | 150K | Yes |
| 13 | No | Married | 80K | Yes |
| 14 | No | Single | 40K | No |
| 15 | No | Married | 80K | Yes |

Trước khi làm ta thấy: dữ liệu dư thừa, trùng nhau nên ta loại bỏ khi đó bảng dữ liệu ta còn như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tid** | **Refund** | **Marital Status** | **Taxable income** | **Cheat** |
| 1 | No | Single | 75K | No |
| 2 | Yes | Married | 50K | No |
| 4 | No | Married | 150K | Yes |
| 5 | No | Single | 40K | No |
| 6 | No | Married | 80K | Yes |

**a, Hãy vẽ cây quyết định và rút ra luật quyết định từ cây quyết định**

Bước 1: Tính Entropy(S)

Bước 2: Tính Entropy của các thuộc tính.

Bước 2.1: Xét thuộc tính Refund có 2 giá trị là Yes, No

Entropy (Refund, S) = Entropy (No)+Entropy (Yes)

= ()+()

= 0,8

Bước 2.2: Xét thuộc tính Marital Status có 2 giá trị Single và Married

Entropy (Marital Status, S) = Entropy (Single)+Entropy (Married)

= ()+()

= 0,551

Bước 2.3: Xét thuộc tính Taxable income có 5 giá trị 40K, 50K, 75K, 80K, 150K

Entropy (Taxable income, S) =Entropy (40K) +Entropy(50K) + Entropy (75K) +Entropy(80K) + Entropy (150K)

Ta có: Entropy(40K) = 0 vì đều có thuộc tính Cheat là No

Etropy(50K) = 0 vì đều có thuộc tính Cheat là No

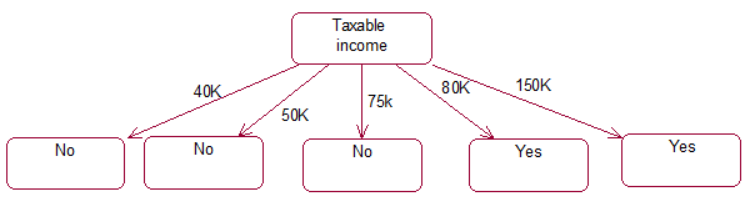
Entropy(75K) = 0 vì đều có thuộc tính Cheat là No

Entropy(80K) = 0 vì đều có thuộc tính Cheat là Yes

Entropy(150K) = 0 vì đều có thuộc tính Cheat là Yes

Nên ta có Entropy (Taxable income, S) = 0

Bước 3: Vì Entropy (Taxable income, S) = 0 là nhỏ nhất nên Gain (Taxable income, S) là lớn nhất) khi đó ta chọn thuộc tính Taxable income làm nút gốc.



Bước 4: Do trường hợp này các thuộc tính đều đã được phân nhánh nên ta không cần tính toán tiếp như vậy ta có luật của cây quyết định là:

* Nếu thuộc tính Taxable income có giá trị là 40K thì thuộc tính Cheat có giá trị là No.
* Nếu thuộc tính Taxable income có giá trị là 50K thì thuộc tính Cheat có giá trị là No.
* Nếu thuộc tính Taxable income có giá trị là 75K thì thuộc tính Cheat có giá trị là No.
* Nếu thuộc tính Taxable income có giá trị là 80K thì thuộc tính Cheat có giá trị là Yes.
* Nếu thuộc tính Taxable income có giá trị là 150K thì thuộc tính Cheat có giá trị là Yes.

**b, Sử dụng cây quyết định đã tạo để dự đoán sau:**

|  |  |  |  |
| --- | --- | --- | --- |
| Refund | Marital Status | Taxable income | Cheat |
| No | Married | 80K | ? |

* Đầu tiên ta xét thuộc tính taxable income có giá trị 80K, vậy theo luật của cây quyết định ta được giá trị Cheat là Yes

|  |  |  |  |
| --- | --- | --- | --- |
| Refund | Marital Status | Taxable income | Cheat |
| No | Married | 80K | **Yes** |

**Bài 2: Ta sử dụng bảng dữ liệu của bài 1 sau khi đã tiền xử lý:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tid** | **Refund** | **Marital Status** | **Taxable income** | **Cheat** |
| 1 | No | Single | 75K | No |
| 2 | Yes | Married | 50K | No |
| 4 | No | Married | 150K | Yes |
| 5 | No | Single | 40K | No |
| 6 | No | Married | 80K | Yes |

**Hãy xây dựng bộ phân lớp Bayes dựa vào bảng mẫu sau và dự đoán mẫu test**

Bước 1: Ta có P (Cheat =Yes) = P (Cheate=No) =

Bước 2: Ta có:

|  |  |  |
| --- | --- | --- |
| **Refund** | Cheat= Yes | Cheat= No |
| Yes | 0/2 | 1/3 |
| No | 2/2 | 2/3 |

|  |  |  |
| --- | --- | --- |
| **Marital Status** | Cheat= Yes | Cheat= No |
| Single | 0/2 | 2/3 |
| Married | 2/2 | 1/3 |

|  |  |  |
| --- | --- | --- |
| **Taxable income** | Cheat= Yes | Cheat= No |
| 40K | 0/2 | 1/3 |
| 50K | 0/2 | 1/3 |
| 75K | 0/2 | 1/3 |
| 80K | 1/2 | 0/3 |
| 150K | 1/2 | 0/3 |

Bước 3:

Gọi mẫu quan sát X = (Refund = No, Marital Status = Married, Taxable income = 80K)

P(Yes|X) = P (Refund=No|Cheat=Yes) \* P (Marital Status =Married|Cheat= Yes) \* P (Taxable income = 80K|Cheat= Yes) \* P (Cheat =Yes)

= \* \* \* = 0,2

P(No|X) = P (Refund=No|Cheat=No) \* P (Marital Status =Married|Cheat= No) \* P (Taxable income = 80K|Cheat= No) \* P (Cheat = No)

= \* \* \* = 0

Do P (Yes|X) > P(No|X) nên X thuộc lớp Yes

**Bài 3:　Có bảng dữ liệu sau:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Morning | Master | Grass | W |
| Afternoon | Grand slam | Glay | W |
| Night | Friendly | Hard | W |
| Afternoon | Friendly | Mixed | N |
| Afternoon | Master | Glay | N |
| Afternoon | Grand slam | Grass | W |
| Afternoon | Grand slam | Hard | W |
| Afternoon | Grand slam | Hard | W |
| Morning | Master | Grass | W |
| Afternoon | Grand slam | Glay | N |
| Night | Friendly | Hard | W |
| Night | Master | Mixed | N |
| Afternoon | Master | Glay | N |
| Afternoon | Master | Grass | W |
| Afternoon | Grand slam | Hard | W |
| Afternoon | Grand slam | Glay | W |

**1, Hãy vẽ cây quyết định và rút ra luật từ cây quyết đinh.**

Trước khi làm ta thấy có những dữ liệu trùng nhau và mâu thuẫn khi đó ta loại bỏ chúng đi khi ấy bảng dữ liệu của ta còn như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Afternoon | Friendly | Mixed | N |
| Afternoon | Grand slam | Grass | W |
| Morning | Master | Grass | W |
| Night | Friendly | Hard | W |
| Night | Master | Mixed | N |
| Afternoon | Master | Glay | N |
| Afternoon | Master | Grass | W |
| Afternoon | Grand slam | Hard | W |
| Afternoon | Grand slam | Glay | W |

Bước 1: Tính Entropy(S)

Bước 2: Tính Entropy của các thuộc tính

Bước 2.1: Xét thuộc tính Time có 3 giá trị là Morning, Afternoon, Night

Entropy (Time, S) = Entropy (Morning)+Entropy (Afternoon)+ Entropy (Night) ta thấy:

Entropy (Morming) = 0 vì chỉ có giá trị Outcome là W nên ta có

Entropy (Time, S) = Entropy (Afternoon)+ Entropy (Night)

= .()+.()) = 0,612 + 0,22 = 0,832

Bước 2.2: Xét thuộc tính Match type có 3 giá trị Master, Grand slam, Friendly

Entropy (Match type, S) = Entropy (Master)+Entropy (Grand slam) + Entropy (Friendly)

= .()+.()) +.() = 0,667

Bước 2.3: Xét thuộc tính Court surfacecó 4 giá trị Glay, Hard, Mixed, Grass

Entropy (Court surface, S) = Entropy (Glay) +Entropy (Hard) + Entropy (Mixed) +Entropy (Grass)

Ta có: Entropy (Mixed) = 0 vì đều có thuộc tính Outcome là N

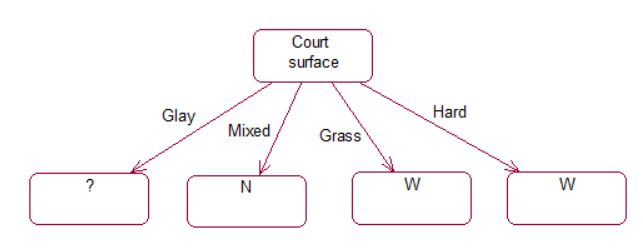
Entropy (Grass) = 0 vì đều có thuộc tính Outcome là W

Entropy (Hard) = 0 vì đều có thuộc tính Outcome là W

Khi đó:

Entropy (Court surface, S) = Entropy (Glay) = (loglog) = 0,22

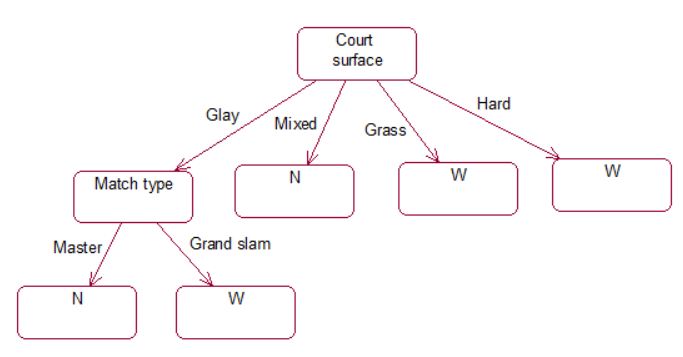
Bước 3: Vì Entropy (Court surface, S) = 0,22 là nhỏ nhất nên Gain (Court surface, S) là lớn nhất) khi đó ta chọn thuộc tính Court surface làm nút gốc.



Bước 4: Khi đã có được một phần cây nhị phân nhưng thuộc tính Glay chưa biết thuộc vào lớp N hay W nên ta tiếp tục tính, khi đó ta còn lại bảng như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Time | Match type | Court surface | Outcome |
| Afternoon | Master | Glay | N |
| Afternoon | Grand slam | Glay | W |

* Đến đây ta chỉ còn 2 thuộc tính chưa được xây dựng là Time với Match type ta thấy một điều thuộc tính Match type có Entropy = 0 vì đối với giá trị là Master chỉ được kết quả Outcome là N và với giá trị Grand slam chỉ được kết quả Outcome là W. như vậy ta chọn được nút con tiếp theo là Match type và thế là ta xây dựng được cây nhị phân như sau:



Khi đó ta rút được ra luật như sau:

* Nếu taxable income có các giá trị là Hard, Mixed thì Outcome sẽ có giá trị là W.
* Nếu taxable income có giá trị là Grass thì Outcome sẽ có giá trị là N.
* Nếu taxable income có giá trị là Glay:
  + Nếu thuộc tính match type có giá trị là Master thì outcome sẽ có giá trị là N
  + Nếu thuộc tính match type có giá trị là Grand slam thì out come sẽ có giá trị là W

**2, Sử dụng cây quyết định đưa ra dự đoán:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Morning | Grand slam | Grass | ? |

* Đầu tiên ta chọn thuộc tính Court surface ta thấy nó có giá trị là Grass, vậy theo luật từ cây quyết định thì giá trị outcome sẽ là N

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Morning | Grand slam | Grass | **W** |

**Bài 4: Ta sử dụng bảng dữ liệu sau khi đã tiền xử lý dữ liệu của bài 3:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Afternoon | Friendly | Mixed | N |
| Afternoon | Grand slam | Grass | W |
| Morning | Master | Grass | W |
| Night | Friendly | Hard | W |
| Night | Master | Mixed | N |
| Afternoon | Master | Glay | N |
| Afternoon | Master | Grass | W |
| Afternoon | Grand slam | Hard | W |
| Afternoon | Grand slam | Glay | W |

Bước 1: Ta có P (Outcome = N) = ; P (Outcome = W) =

Bước 2: Ta có:

|  |  |  |
| --- | --- | --- |
| **Time** | Outcome = W | Outcome = N |
| Afternoon | 4/6 | 2/3 |
| Morning | 1/6 | 0/3 |
| Night | 1/6 | 1/3 |

|  |  |  |
| --- | --- | --- |
| **Match type** | Outcome = W | Outcome = N |
| Friendly | 1/6 | 1/3 |
| Grand slam | 3/6 | 0/3 |
| Master | 2/6 | 2/3 |

|  |  |  |
| --- | --- | --- |
| **Court surface** | Outcome = W | Outcome = N |
| Mixed | 0/6 | 2/3 |
| Glay | 1/6 | 1/3 |
| Grass | 3/6 | 0/3 |
| Hard | 2/6 | 0/3 |

Bước 3:

Gọi mẫu quan sát X = (Time = Morning, Match type = Grand slam, Court surface = Grass)

P(W|X) = P (Time = Morning | Outcome = W) \* P (Match type = Grand slam | Outcome = W) \* P (Court surface = Grass| Outcome = W) \* P (Outcome = W)

= \* \* \* =

P(N|X) = P (Time = Morning | Outcome = N) \* P (Match type = Grand slam | Outcome = N) \* P (Court surface = Grass| Outcome = N) \* P (Outcome = N)

= \* \* \* = 0

Do P(W|X) > P(N|X) nên X thuộc lớp W (Giống kết quả bài 3)