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

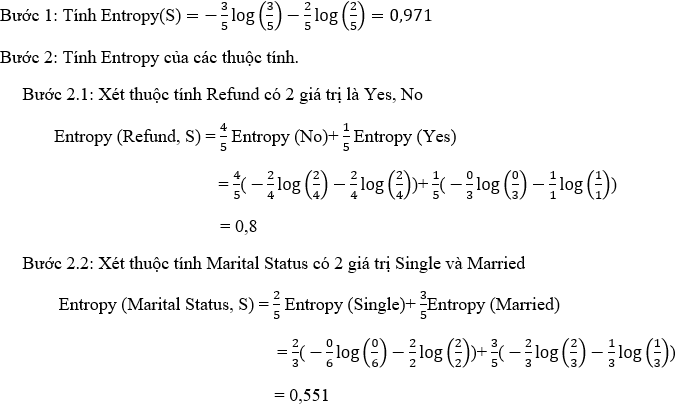
**Bài 1: Cho dữ liệu thu thập được như sau:**

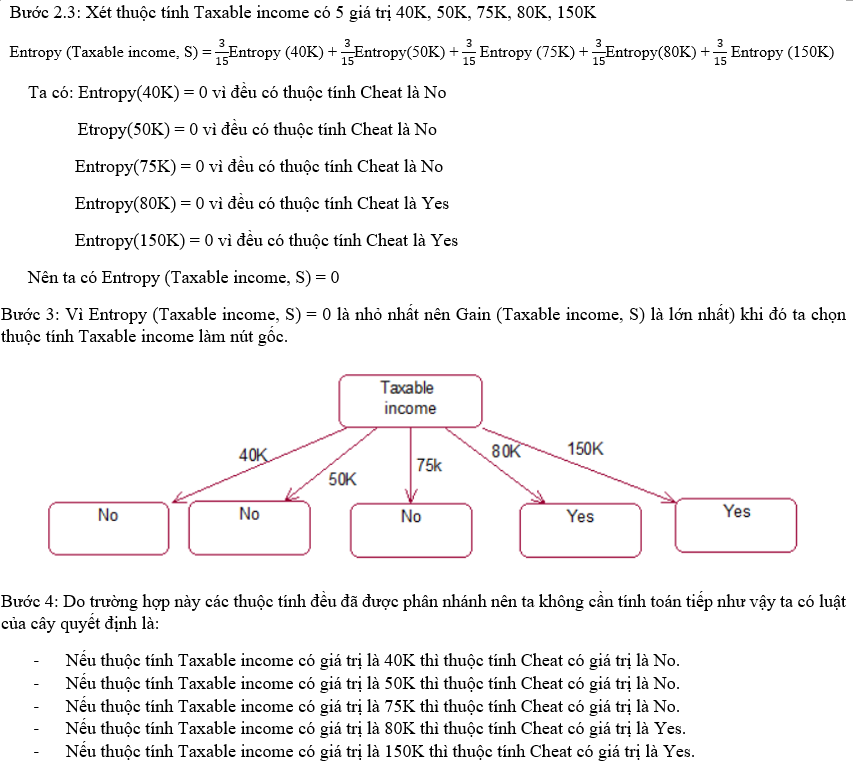
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tid** | **Refund** | **Marital Status** | **Taxable income** | **Cheat** |
| 1 | No | Single | 75K | No |
| 2 | Yes | Married | 50K | No |
| 3 | No | Single | 75K | No |
| 4 | No | Married | 150K | Yes |
| 5 | No | Single | 40K | No |
| 6 | No | Married | 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**





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

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

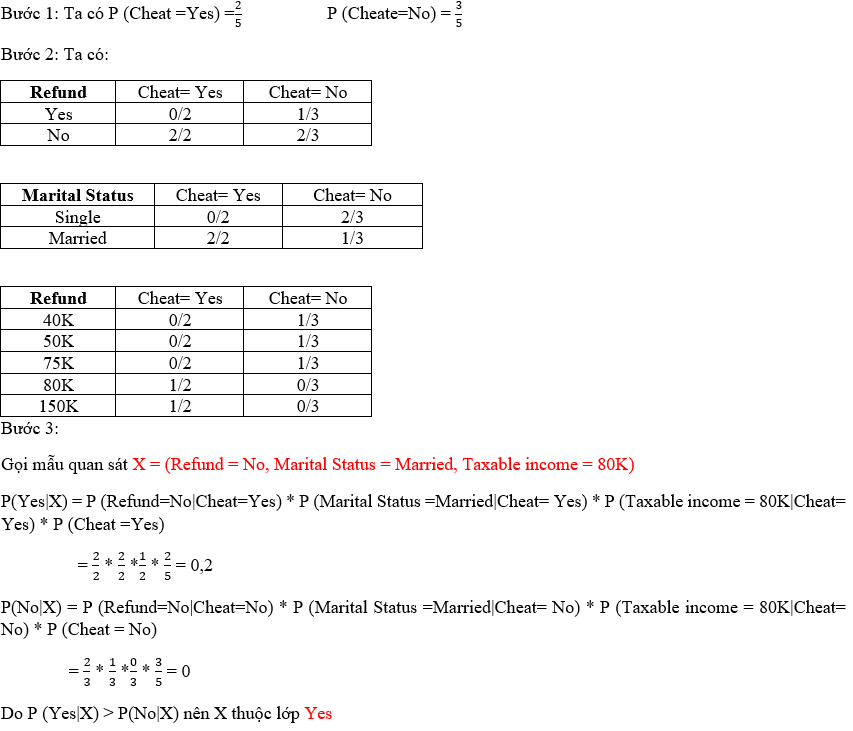
xét thuộc tính taxable income có giá trị 80K, vậy theo luật của cây quyết định ta được giá trị Cheat là Yes

|  |  |  |  |
| --- | --- | --- | --- |
| 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**



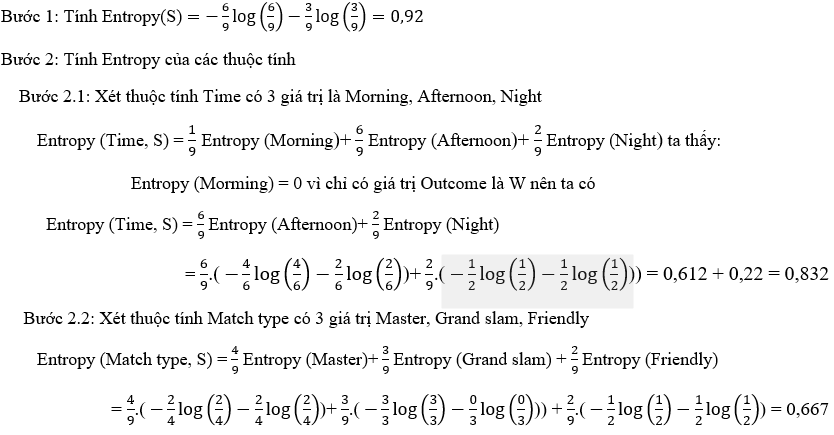
**Bài 3:　Có bảng dữ liệu sau:**

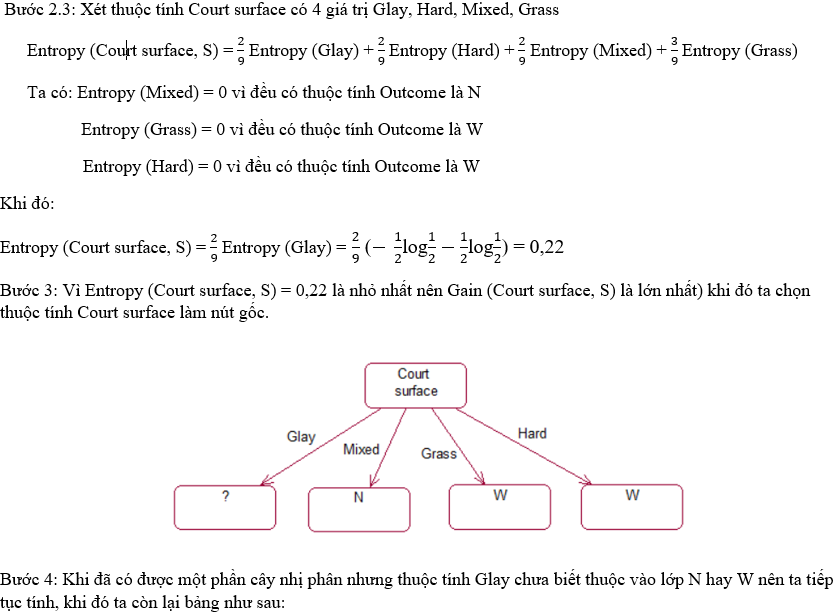
|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Morning | Master | Grass | W |
| Afternoon | Grand slam | Glay | W |
| Night | Friendly | Hard | W |
| Afternoon | Friendly | Mixed | N |
| Afternoon | Master | Glay | N |
| Afternoon | Grand slam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

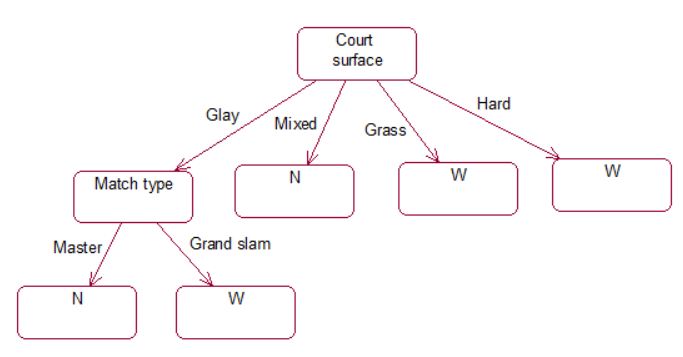
|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Match type** | **Court surface** | **Outcome** |
| Afternoon | Friendly | Mixed | N |
| Afternoon | Grand slam | Grass | W |
| Morning | Master | Grass | W |
| Night | Friendly | Hard | W |
| Night | Master | Mixed | N |
| Afternoon | Master | Glay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





|  |  |  |  |
| --- | --- | --- | --- |
| 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 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 |

