BÀI TẬP KHAI PHÁ DỮ LIỆU

**Vẽ cây quyết định**

Theo bảng dữ liệu ta có:

- |D| = 8

* C = {Bị rám, không}
* |CBị rám| = 3, |CKhông| = 5 Tính Entropy
* Info (D) =

Xét thuộc tính màu tóc:

|  |  |  |
| --- | --- | --- |
| Đen | Bị rám | 2 |
| Không | 2 |
| Rám | Bị rám | 0 |
| Không | 3 |
| Bạc | Bị rám | 1 |
| Không | 0 |

Infomàu tóc (D) =

 Gain (màu tóc) = 0.954bits – 0.5bits = 0.454bits

Xét thuộc tính chiều cao:

|  |  |  |
| --- | --- | --- |
| Cao | Bị rám | 0 |
| Không | 2 |
| Tầm thước | Bị rám | 2 |
| Không | 1 |
| Thấp | Bị rám | 1 |
| Không | 2 |

Infochiều cao (D) =

 Gain (chiều cao) = 0.954bits – 0.689bits = 0.265bits

Xét thuộc tính cân nặng:

|  |  |  |
| --- | --- | --- |
| Nhẹ | Bị rám | 1 |
| Không | 1 |
| Vừa phải | Bị rám | 1 |
| Không | 2 |
| Nặng | Bị rám | 1 |
| Không | 2 |

Infocân nặng (D) =

 Gain (cân nặng) = 0.954bits – 0.939bits = 0.015bits

Xét thuộc tính dùng thuốc:

|  |  |  |
| --- | --- | --- |
| Không | Bị rám | 3 |
| Không | 2 |
| Có | Bị rám | 0 |
| Không | 3 |

Infodùng thuốc (D) = ts

 Gain (dùng thuốc) = 0.954bits – 0.607bits = 0.347bit

Vì thuộc tính *màu tóc* có độ lợi thông tin lớn nhất (0.454bits) nên *màu tóc* là thuộc tính được chọn để phân tách (rẽ nhánh)

Ta có cây quyết định:

Màu tóc

Đen

Bạc

Rám

Dùng thuốc

Không

Bị rám

Không

Có

Bị rám

Không