**Đề:** Using ID3 algorithm, construct (by hand, show detail work) a decision tree for below dataset. Is your decision tree overfitting? Can you use Pruning technique to reduce the problem? Explain your work.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Example** | **mpg** | **cylinders** | **displacement** | **horsepower** | **weight** | **acceleration** | **modelyear** | **maker** |
| 1 | good | 4 | low | low | low | high | 75to78 | asia |
| 2 | bad | 6 | medium | medium | medium | medium | 70to74 | america |
| 3 | bad | 4 | medium | medium | medium | low | 75to78 | europe |
| 4 | bad | 8 | high | high | high | low | 70to74 | america |
| 5 | bad | 6 | medium | medium | medium | medium | 70to74 | america |
| 6 | bad | 4 | low | medium | low | medium | 70to74 | asia |
| 7 | bad | 4 | low | medium | low | low | 70to74 | asia |
| 8 | bad | 8 | high | high | high | low | 75to78 | america |
| 9 | bad | 8 | high | high | high | low | 70to74 | america |
| 10 | good | 8 | high | medium | high | high | 79to83 | america |
| 11 | bad | 8 | high | high | high | low | 75to78 | america |
| 12 | good | 4 | low | low | low | low | 79to83 | america |
| 13 | bad | 6 | medium | medium | medium | high | 75to78 | america |
| 14 | good | 4 | medium | low | low | low | 79to83 | america |
| 15 | good | 4 | low | low | medium | high | 79to83 | america |
| 16 | bad | 8 | high | high | high | low | 70to74 | america |
| 17 | good | 4 | low | medium | low | medium | 75to78 | europe |
| 18 | bad | 5 | medium | medium | medium | medium | 75to78 | europe |

**Predict Miles-per-gallon?**

**Ký hiệu:** +(good), - (bad)

**Lặp lần 1:**

Entropy(S) = = = 0.5283 + 0.39 = 0.9183

|  |  |
| --- | --- |
| **cylinders** | |
| **4:5+,3-** | Entropy(S4) = = 0.9544. |
| **5:0+,1-** | Entropy(S5) = = 0. |
| **6:0+,3-** | Entropy(S6) = = 0. |
| **8:1+,5-** | Entropy(S8) = = 0.650. |
| **Gain(S,cylinders)** =)  =  = = 0.2774. | |
| **displacement** | |
| **high:1+,5-** | Entropy(Shigh) = = 0.650 |
| **Medium:1+,5-** | Entropy(Smedium) = = 0.650 |
| **Low:4+,2-** | Entropy(Slow) = = 0.9183. |
| **Gain(S, displacement)** = )  = = 0.1789. | |

|  |  |
| --- | --- |
| **horsepower** | |
| **high:0+,5-** | Entropy(Shigh) = = 0 |
| **Medium:2+,7-** | Entropy(Smedium) = = 0.7642 |
| **Low:4+,0-** | Entropy(Slow) = = 0 |
| **Gain(S, horsepower)** = )  = = 0.5362. | |

|  |  |
| --- | --- |
| **weight** | |
| **high:1+,5-** | Entropy(Shigh) = = 0.650 |
| **Medium:1+,5-** | Entropy(Smedium) = = 0.650 |
| **Low:4+,2-** | Entropy(Slow) = = 0.9183. |
| **Gain(S, weight)** = )  = = 0.1789 | |

|  |  |
| --- | --- |
| **acceleration** | |
| **high:3+,1-** | Entropy(Shigh) = = 0.8113 |
| **Medium:1+,4-** | Entropy(Smedium) = = 0.7219 |
| **Low:2+,7-** | Entropy(Slow) = =0 .7642. |
| **Gain(S, acceleration)** = )  = = 0.1554 | |

|  |  |
| --- | --- |
| **modelyear** | |
| **70to74:0+,7-** | Entropy(S70to74) = = 0 |
| **75to78:2+,5-** | Entropy(S75to78) = = 0.8631 |
| **79to83:4+,0-** | Entropy(S79to83) = = 0 |
| **Gain(S, modelyear)** = )  = = 0.5826. | |

|  |  |
| --- | --- |
| **maker** | |
| **asia:1+,2-** | Entropy(Sasia) = = 0.9183 |
| **america:4+,8-** | Entropy(Samerica) = = 0.9183 |
| **europe:1+,2-** | Entropy(Seurope) = = 0.9183 |
| **Gain(S, modelyear)** = ) = | |

**Tổng hợp kết quả lặp lần 1:**

|  |  |  |
| --- | --- | --- |
| **STT** | **Name** | **Gain** |
| 1 | cylinders | 0.2774 |
| 2 | displacement | 0.1789 |
| 3 | horsepower | 0.5362 |
| 4 | weight | 0.1789 |
| 5 | acceleration | 0.1554 |
| 6 | modelyear | 0.5826 |
| 7 | maker | 0 |

**Dựa trên kết quả Gain có giá trị càng lớn càng tốt. Vì vậy ta chọn " horsepower "**

**modelyear**

**79to83**

**75to78**

**?**

good

bad

**70to74**

**Lặp lần 2:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Example** | **mpg** | **cylinders** | **displacement** | **horsepower** | **weight** | **acceleration** | **maker** |
| 1 | good | 4 | low | low | low | high | asia |
| 2 | bad | 4 | medium | medium | medium | low | europe |
| 3 | bad | 8 | high | high | high | low | america |
| 4 | bad | 8 | high | high | high | low | america |
| 5 | bad | 6 | medium | medium | medium | high | america |
| 6 | good | 4 | low | medium | low | medium | europe |
| 7 | bad | 5 | medium | medium | medium | medium | europe |

Entropy(S) = = = 0.5164 + 0.3467 = 0.8631

|  |  |
| --- | --- |
| **cylinders** | |
| **4:2+,1-** | Entropy(S4) = = 0.9183 |
| **5:0+,1-** | Entropy(S5) = = 0. |
| **6:0+,1-** | Entropy(S6) = = 0. |
| **8:0+,2-** | Entropy(S8) = = 0. |
| **Gain(S,cylinders)** =)  =  = = 0.4696. | |

|  |  |
| --- | --- |
| **displacement** | |
| **high:0+,2-** | Entropy(Shigh) = = 0 |
| **Medium:0+,3-** | Entropy(Smedium) = = 0 |
| **Low:2+,0-** | Entropy(Slow) = = 0 |
| Gain(S, displacement) = )  = = 0.8631 | |

|  |  |
| --- | --- |
| **horsepower** | |
| **high:0+,2-** | Entropy(Shigh) = = 0 |
| **Medium:1+,3-** | Entropy(Smedium) = = 0.8113 |
| **Low:1+,0-** | Entropy(Slow) = = 0 |
| Gain(S, horsepower) = )  = = 0.3995 | |

|  |  |
| --- | --- |
| **weight** | |
| **high:0+,2-** | Entropy(Shigh) = = 0 |
| **Medium:0+,3-** | Entropy(Smedium) = = 0 |
| **Low:2+,0-** | Entropy(Slow) = = 0 |
| **Gain(S, weight)** = )  = = 0.8631 | |

|  |  |
| --- | --- |
| **acceleration** | |
| **high:1+,1-** | Entropy(Shigh) = = 1 |
| **Medium:1+,1-** | Entropy(Smedium) = = 1 |
| **Low:0+,3-** | Entropy(Slow) = = 0 |
| **Gain(S, acceleration)** = )  = = 0.2917 | |

|  |  |
| --- | --- |
| **maker** | |
| **asia:1+,0-** | Entropy(Sasia) = = 0 |
| **america:0+,3-** | Entropy(Samerica) = = 0 |
| **europe:1+,2-** | Entropy(Seurope) = = 0.9183 |
| **Gain(S, maker)** = )  = = 0.4696. | |

**Tổng hợp kết quả lặp lần 2:**

|  |  |  |
| --- | --- | --- |
| **STT** | **Name** | **Gain** |
| 1 | cylinders | 0.4696 |
| 2 | displacement | 0.8631 |
| 3 | horsepower | 0.3995 |
| 4 | weight | 0.8631 |
| 5 | acceleration | 0.2917 |
| 6 | maker | 0.4696 |

**Dựa trên kết quả Gain có giá trị càng lớn càng tốt. Nhưng dựa vào bảng kết quả ta thấy có 2 giá trị bằng nhau “displacement”, “weight”. Vì vậy ta có thể chọn 1 trong 2 " displacement "**

**modelyear**

**79to83**

**75to78**

**displacement**

good

bad

**70to74**

**high**

**medium**

**low**

bad

good

bad

**Kết luận:** Mô hình cây quyết định đơn giản, có thể áp dụng cho bải toán trên.

* if modelyear = 70to74 then bad
* if modelyear = 79to83 then good
* if modelyear = 75to78 and displacement = low then good
* if modelyear = 75to78 and displacement = medium then bad
* if modelyear = 75to78 and displacement = high then bad