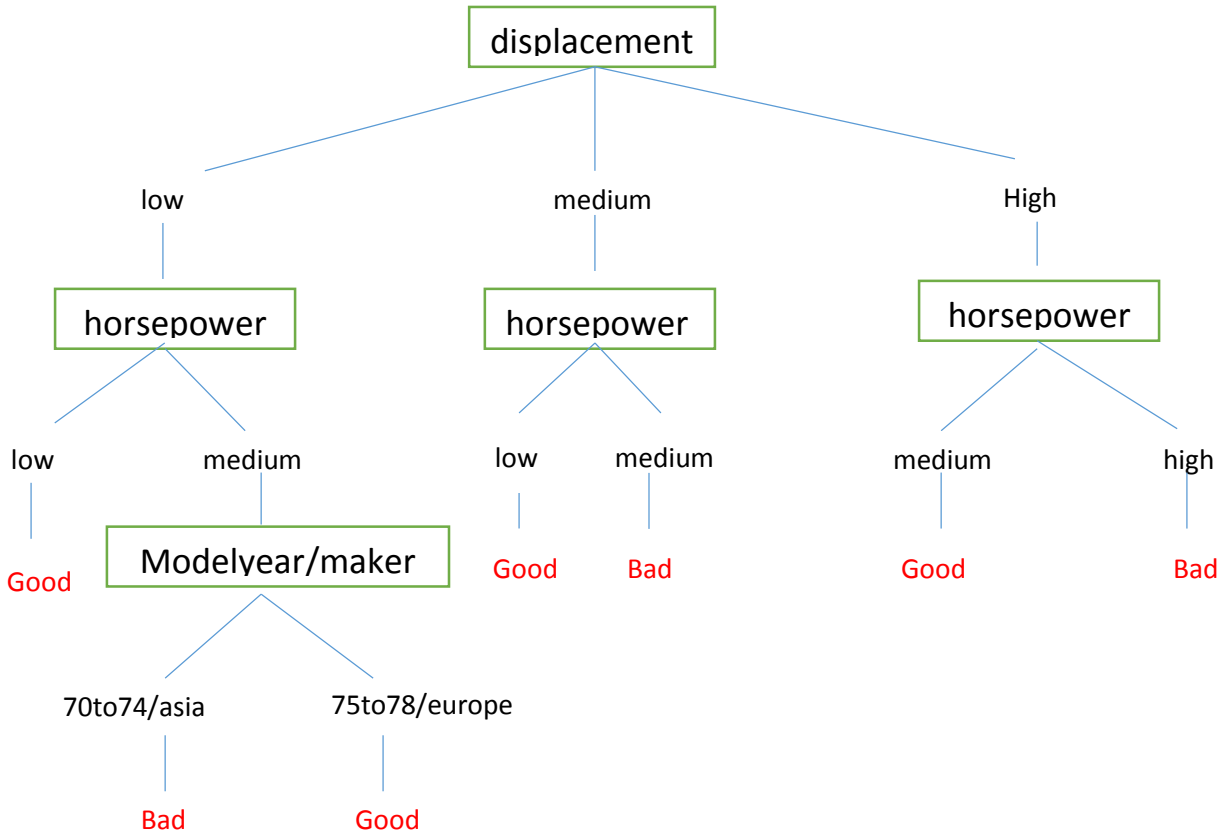
Decision Trees

- **Case 1:** Nếu chọn bắt đầu '**cylinders**' thì với giá trị cylinders = 4: tiếp tục đến '**horsepower**' với medium -> '**acceleration**' với medium -> '**modelyear**' mới đến được '**pmg**'. Như vậy cần 4 cấp feature.
- **Case 2**: Bắt đầu với '**displacement**'.



Tính mức độ hỗn loạn các phân tử trong đối tượng: Với good = 6, bad = 12.

Entropy(S) = - p+log2p+ - p-log2p- = - (6/18)log2(6/18) - (12/18)log2(12/18) = 0.5261 + 0.3896 = 0.9157

Để xác định feature nào là tốt nhất tính:

Gain(S, displacement)
= Entropy(S) - ∑v∈{ displacement-low, displacement-medium, displacement-high }(|Sv|/|S|)Entropy(Sv)

= Entropy(S) – [(6/18).Entropy(S displacement-low) + (6/18).Entropy(S displacement-medium) + (6/18).Entropy(S displacement-high)]

= Entropy(S) – [(6/18).Entropy(S displacement-low) + 0 + 0]

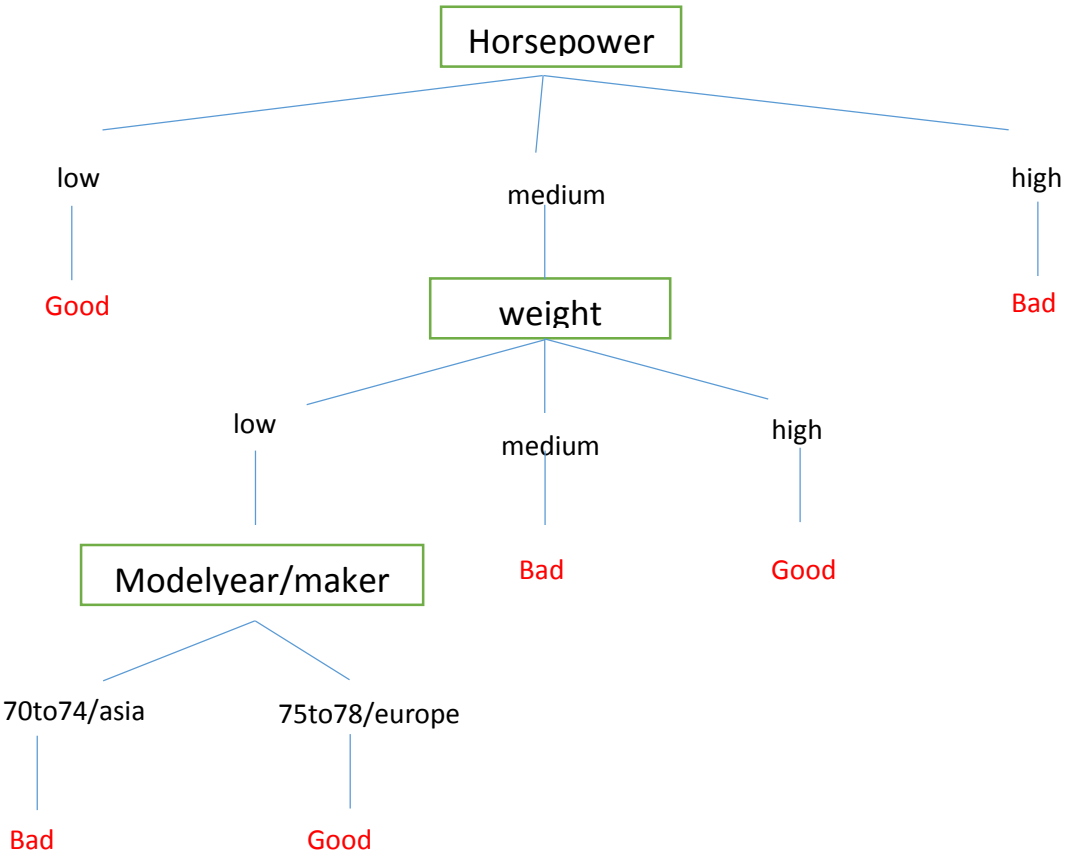
= Entropy(S) – [(6/18).[(3/6)Entropy(S horsepower-low) + (3/6)Entropy(S horsepower-medium)]]

= Entropy(S) – [(6/18).[0 + (3/6)Entropy(S horsepower-medium)]

= Entropy(S) – [(6/18). (3/6) [Entropy(S Modelyear -70to74) + Entropy(S Modelyear -75to78)]]

= 0.9157 – (6/18)(3/6)(- (1/2)log2(1/2) - (1/2)log2(1/2)) = 0.9157 – 0.1667 = 0.794

Case 3: Bắt đầu với '**Horsepower**'



Gain(S, horsepower) = Entropy(S) - ∑v∈{ horsepower -low, horsepower -medium, horsepower -high }(|Sv|/|S|)Entropy(Sv)

= Entropy(S) – [(4/18).Entropy(S horsepower -low) + (9/18).Entropy(S horsepower -medium) + (5/18).Entropy(S horsepower -high)]

= Entropy(S) – [0 + (5/18).Entropy(S horsepower - medium) + 0]

= Entropy(S) – [(5/18).[(3/9) Entropy(S weight -low) + (5/9) Entropy(S weight - medium) + (1/9) Entropy(S weight - high)]]

= Entropy(S) – [(5/18).[(3/9) Entropy(S weight -low) + 0 + 0]]

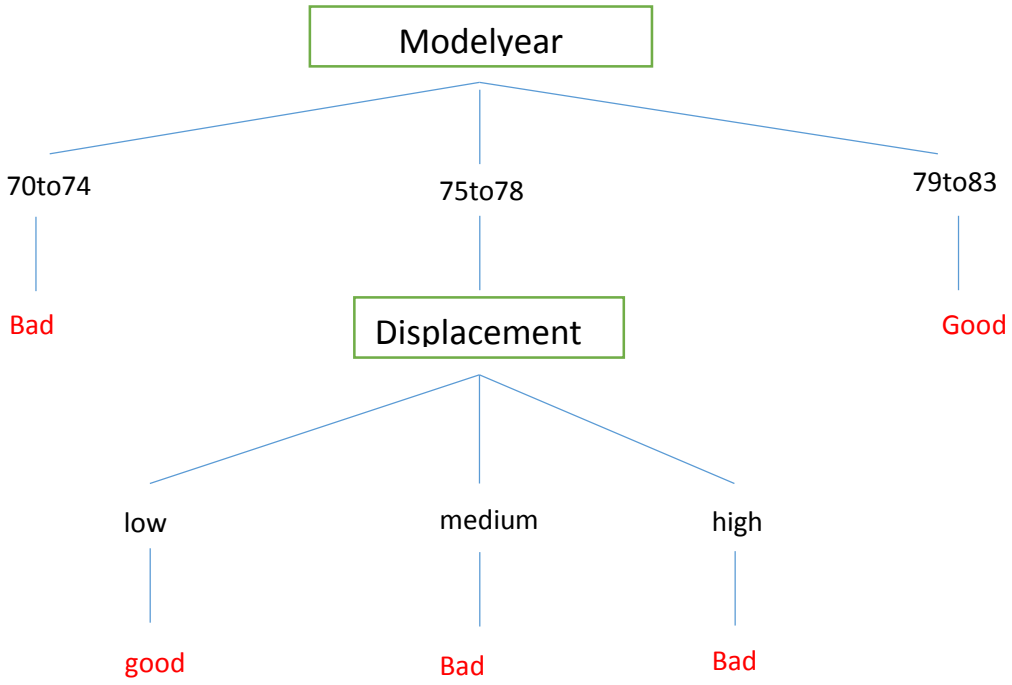
= 0.9157 – (5/18)(3/9)(- (1/2)log2(1/2) - (1/2)log2(1/2))

= 0.9157 – 0.0922 = 0.8235

Case 4: Nếu chọn bắt đầu từ '**weight**' -> '**horsepower**' thì sẽ lặp lại giống như case 2.

Case 5: Nếu chọn bắt đầu từ '**acceleration**' với low -> '**deplacement**' với low -> '**horsepower**'/'**modelyear**' -> 'pmg' cần 3 cấp feature giống với case 3.

Case 6: modelyear -> 'weight' hoặc 'placement'



Gain(S, Modelyear)
= Entropy(S) - ∑v∈{ Modelyear -70to74, Modelyear -75to78, Modelyear -79to83 }(|Sv|/|S|)Entropy(Sv)

= Entropy(S) – [(7/18).Entropy(S Modelyear -70to74) + (7/18).Entropy(S Modelyear -75to78) + (4/18).Entropy(S Modelyear -75to78)]

= Entropy(S) – [0 + (7/18).Entropy(S Modelyear -75to78) + 0]

= Entropy(S) – [(7/18).Entropy(S displacement)]

= Entropy(S) – [(7/18).[(3/7) Entropy(S displacement -low) + (2/7).Entropy(S displacement -medium) + (2/7).Entropy(S displacement -high)]]

= 0.9157 – 0.3889 = 0.5268

Case 7: '**maker**' với america -> '**weight**' với high -> '**horsepower**' -> '**pmg**'. Cần 3 cấp feature để đến 'pmg'.

Nhận xét: qua giá trị Gain(S,feature) ta thấy case 6 với modelyear là Gain(S, Modelyear) nhỏ nhất nên tốt nhất.