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) = $-P_+log_2(P_+) - P\_log_2(P\_) = -\frac{6}{18}log_2(\frac{6}{18}) - \frac{12}{18}log_2(\frac{12}{18})$ = 0.5283 + 0.39 = 0.9183

cylinders	
4:5+,3-	Entropy(S₄) = $[-\frac{5}{8}log_2(\frac{5}{8}) - \frac{3}{8}log_2(\frac{3}{8})]$ = 0.9544.
5:0+,1-	Entropy(S₅) = $[-\frac{1}{1}log_2(\frac{1}{1})]$ = 0.
6:0+,3-	Entropy(S₆) = $[-\frac{3}{3}log_2(\frac{3}{3})]$ = 0.
8:1+,5-	Entropy(S₈) = $[-\frac{1}{6}log_2(\frac{1}{6}) - \frac{5}{6}log_2(\frac{5}{6})]$ = 0.650.
Gain(S,cylinders) = $Entropy(S) - \sum_{v\in\{4,5,6,8\}}(\lvert S_v\rvert/\lvert S\rvert)Entropy(S_v)$ $= Entropy(S) - (\frac{8}{18} * Entropy(S_4) + \frac{1}{18} * Entropy(S_5) + \frac{3}{18} * Entropy(S_6) + \frac{6}{18} * Entropy(S_8))$ $= 0.9183 - (\frac{8}{18} * 0.9544 + \frac{1}{18} * 0 + \frac{3}{18} * 0 + \frac{6}{18} * 0.650)$ = 0.2774.	

<table>
<tr><th colspan="2">displacement</th></tr>
<tr><td><b>high:1+,5-</b></td><td>Entropy(S_{high}) = $[-\frac{1}{6}log_2(\frac{1}{6}) - \frac{5}{6}log_2(\frac{5}{6})]$ = 0.650</td></tr>
<tr><td><b>Medium:1+,5-</b></td><td>Entropy(S_{medium}) = $[-\frac{1}{6}log_2(\frac{1}{6})-\frac{5}{6}log_2(\frac{5}{6})]$ = 0.650</td></tr>
<tr><td><b>Low:4+,2-</b></td><td>Entropy(S_{low}) = $[-\frac{4}{6}log_2(\frac{4}{6})-\frac{2}{6}log_2(\frac{2}{6})]$ = 0.9183.</td></tr>
<tr><td colspan="2"><b>Gain(S, displacement)</b> = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(|S_v|/|S|)Entropy(S_v)$
= 0.9183 $-(\frac{6}{18}*0.650+\frac{6}{18}*0.650+\frac{6}{18}*0.9183)$ = 0.1789.</td></tr>
</table>

<table>
<tr><th colspan="2">horsepower</th></tr>
<tr><td><b>high:0+,5-</b></td><td>Entropy(S_{high}) = $[-\frac{5}{5}log_2(\frac{5}{5})]$ = 0</td></tr>
<tr><td><b>Medium:2+,7-</b></td><td>Entropy(S_{medium}) = $[-\frac{2}{9}log_2(\frac{2}{9})-\frac{7}{9}log_2(\frac{7}{9})]$ = 0.7642</td></tr>
<tr><td><b>Low:4+,0-</b></td><td>Entropy(S_{low}) = $[-\frac{4}{4}log_2(\frac{4}{4})]$ = 0</td></tr>
<tr><td colspan="2"><b>Gain(S, horsepower)</b> = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(|S_v|/|S|)Entropy(S_v)$
= $0.9183-(\frac{5}{18}*0+\frac{9}{18}*0.5671+\frac{4}{18}*0)$ = 0.5362.</td></tr>
</table>

<table>
<tr><th colspan="2">weight</th></tr>
<tr><td><b>high:1+,5-</b></td><td>Entropy(S_{high}) = $[-\frac{1}{6}log_2(\frac{1}{6}) - \frac{5}{6}log_2(\frac{5}{6})]$ = 0.650</td></tr>
<tr><td><b>Medium:1+,5-</b></td><td>Entropy(S_{medium}) = $[-\frac{1}{6}log_2(\frac{1}{6}) - \frac{5}{6}log_2(\frac{5}{6})]$ = 0.650</td></tr>
<tr><td><b>Low:4+,2-</b></td><td>Entropy(S_{low}) = $[-\frac{4}{6}log_2(\frac{4}{6})-\frac{2}{6}log_2(\frac{2}{6})]$ = 0.9183.</td></tr>
<tr><td colspan="2"><b>Gain(S, weight)</b> = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(|S_v|/|S|)Entropy(S_v)$
= $0.9183-(\frac{6}{18}*0.650+\frac{6}{18}*0.650+\frac{6}{18}*0.9183)$ = 0.1789</td></tr>
</table>

acceleration	
high:3+,1-	Entropy(S_{high}) $= [-\frac{3}{4}log_2(\frac{3}{4})-\frac{1}{4}log_2(\frac{1}{4})]$ = 0.8113
Medium:1+,4-	Entropy(S_{medium}) $= [-\frac{1}{5}log_2(\frac{1}{5})-\frac{4}{5}log_2(\frac{4}{5})]$ = 0.7219
Low:2+,7-	Entropy(S_{low}) $= [-\frac{2}{9}log_2(\frac{2}{9})-\frac{7}{9}log_2(\frac{7}{9})]$ =0 .7642.
Gain(S, acceleration) $= Entropy(S) - \sum_{v\in\{high,medium,low\}}(\lvert S_v\rvert/\lvert S\rvert)Entropy(S_v)$ $= 0.9183 - (\frac{4}{18} * 0.8113 + \frac{5}{18} * 0.7219 + \frac{9}{18} * 0\ .7642) = 0.1554$	

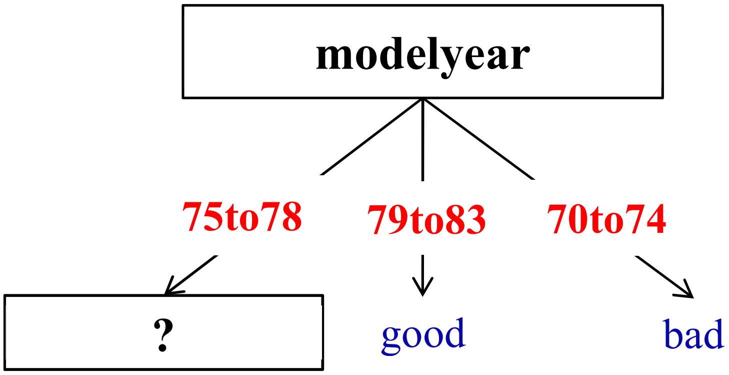
modelyear	
70to74:0+,7-	Entropy(S_{70to74}) $= [-\frac{7}{7}log_2(\frac{7}{7})]$ = 0
75to78:2+,5-	Entropy(S_{75to78}) $= [-\frac{2}{7}log_2(\frac{2}{7})-\frac{5}{7}log_2(\frac{5}{7})]$ = 0.8631
79to83:4+,0-	Entropy(S_{79to83}) $= [-\frac{4}{4}log_2(\frac{4}{4})$= 0
Gain(S, modelyear) $= Entropy(S) - \sum_{v\in\{70to74,75to78,79to83\}}(\lvert S_v\rvert/\lvert S\rvert)Entropy(S_v)$ $= 0.9183 - (\frac{7}{18} * 0 + \frac{7}{18} * 0.8631 + \frac{4}{18} * 0) = 0.5826.$	

maker	
asia:1+,2-	Entropy(S_{asia}) $= [-\frac{1}{3}log_2(\frac{1}{3})-\frac{2}{3}log_2(\frac{2}{3})]$ = 0.9183
america:4+,8-	Entropy($S_{america}$) $= [-\frac{4}{12}log_2(\frac{4}{12})-\frac{8}{12}log_2(\frac{8}{12})]$ = 0.9183
europe:1+,2-	Entropy(S_{europe}) $= [-\frac{1}{3}log_2(\frac{1}{3})-\frac{2}{3}log_2(\frac{2}{3})]$ = 0.9183
Gain(S, modelyear) $= Entropy(S) - \sum_{v\in\{70to74,75to78,79to83\}}(\lvert S_v\rvert/\lvert S\rvert)Entropy(S_v) = 0.9183 - (\frac{3}{18} * 0.9183 + \frac{12}{18} * 0.9183 + \frac{3}{18} * 0.9183) = 0$	

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 "



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) = $-P_{+}log_2(P_{+}) - P_{-}log_2(P_{-}) = -\frac{2}{7}log_2(\frac{2}{7}) - \frac{5}{7}log_2(\frac{5}{7})$ = 0.5164 + 0.3467 = 0.8631

cylinders	
4:2+,1-	Entropy(S_4) = $[-\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3})]$ = 0.9183
5:0+,1-	Entropy(S_5) = $[-\frac{1}{1}log_2(\frac{1}{1})]$ = 0.
6:0+,1-	Entropy(S_6) = $[-\frac{1}{1}log_2(\frac{1}{1})]$ = 0.
8:0+,2-	Entropy(S_8) = $[-\frac{2}{2}log_2(\frac{2}{2})]$ = 0.
Gain(S,cylinders) = $Entropy(S) - \sum_{v\in\{4,5,6,8\}}(\|S_v\|/\|S\|)Entropy(S_v)$ = $Entropy(S) - (\frac{4}{9} * Entropy(S_4) + \frac{1}{9} * Entropy(S_5) + \frac{3}{9} * Entropy(S_6) + \frac{1}{9} * Entropy(S_8))$ = $0.8631 - (\frac{3}{7} * 0.9183 + \frac{1}{7} * 0 + \frac{1}{7} * 0 + \frac{2}{7} * 0)$ = 0.4696.	

displacement	
high:0+,2-	Entropy(S_{high}) = $[-\frac{2}{2}log_2(\frac{2}{2})]$ = 0
Medium:0+,3-	Entropy(S_{medium}) = $[-\frac{3}{3}log_2(\frac{3}{3})]$ = 0
Low:2+,0-	Entropy(S_{low}) = $[-\frac{2}{2}log_2(\frac{2}{2})]$ = 0
Gain(S, displacement) = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(\|S_v\|/\|S\|)Entropy(S_v)$ = $0.8631 - (\frac{2}{7} * 0 + \frac{3}{7} * 0 + \frac{2}{7} * 0)$ = 0.8631	

<table>
<tr><th colspan="2">horsepower</th></tr>
<tr><td><b>high:0+,2-</b></td><td>Entropy(S_{high}) = $[-\frac{2}{2}log_2(\frac{2}{2})] = 0$</td></tr>
<tr><td><b>Medium:1+,3-</b></td><td>Entropy(S_{medium}) = $[-\frac{1}{4}log_2(\frac{1}{4}) - \frac{3}{4}log_2(\frac{3}{4})] = 0.8113$</td></tr>
<tr><td><b>Low:1+,0-</b></td><td>Entropy(S_{low}) = $[-\frac{1}{1}log_2(\frac{1}{1})] = 0$</td></tr>
<tr><td colspan="2">Gain(S, horsepower) = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(|S_v|/|S|)Entropy(S_v)$
$= 0.8631 - (\frac{2}{7} * 0 + \frac{4}{7} * 0.8113 + \frac{1}{7} * 0) = 0.3995$</td></tr>
</table>

<table>
<tr><th colspan="2">weight</th></tr>
<tr><td><b>high:0+,2-</b></td><td>Entropy(S_{high}) = $[-\frac{2}{2}log_2(\frac{2}{2})] = 0$</td></tr>
<tr><td><b>Medium:0+,3-</b></td><td>Entropy(S_{medium}) = $[-\frac{3}{3}log_2(\frac{3}{3})] = 0$</td></tr>
<tr><td><b>Low:2+,0-</b></td><td>Entropy(S_{low}) = $[-\frac{2}{2}log_2(\frac{2}{2})] = 0$</td></tr>
<tr><td colspan="2"><b>Gain(S, weight)</b> = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(|S_v|/|S|)Entropy(S_v)$
$= 0.8631 - (\frac{2}{7} * 0 + \frac{3}{7} * 0 + \frac{2}{7} * 0) = 0.8631$</td></tr>
</table>

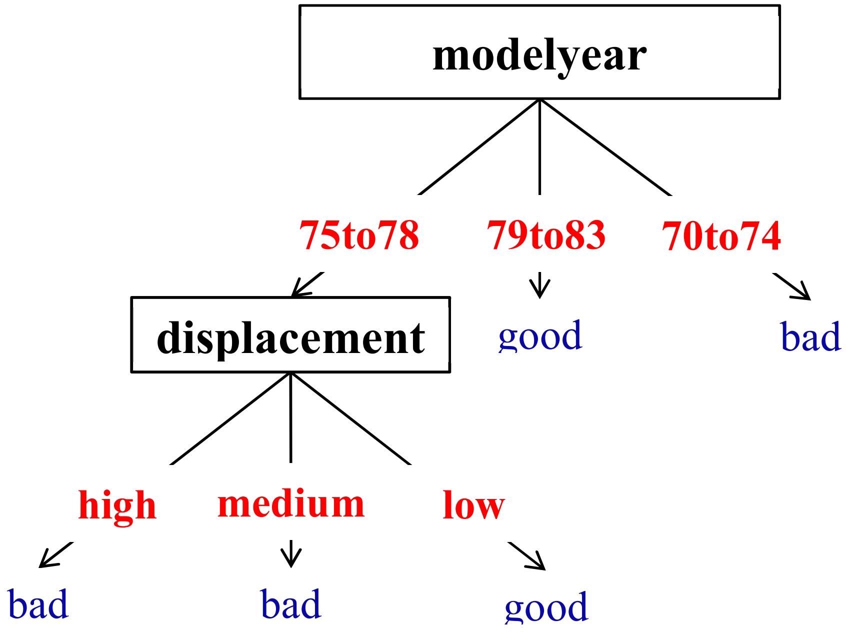
<table>
<tr><th colspan="2">acceleration</th></tr>
<tr><td><b>high:1+,1-</b></td><td>Entropy(S_{high}) = $[-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2})] = 1$</td></tr>
<tr><td><b>Medium:1+,1-</b></td><td>Entropy(S_{medium}) = $[-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2})] = 1$</td></tr>
<tr><td><b>Low:0+,3-</b></td><td>Entropy(S_{low}) = $[-\frac{3}{3}log_2(\frac{3}{3})] = 0$</td></tr>
<tr><td colspan="2"><b>Gain(S, acceleration)</b> = $Entropy(S) - \sum_{v\in\{high,medium,low\}}(|S_v|/|S|)Entropy(S_v)$
$= 0.8631 - (\frac{2}{7} * 1 + \frac{2}{7} * 1 + \frac{3}{7} * 0) = 0.2917$</td></tr>
</table>

maker	
asia:1+,0-	Entropy(S_{asia}) = $[-\frac{1}{1}log_2(\frac{1}{1})] = 0$
america:0+,3-	Entropy($S_{america}$) = $[-\frac{3}{3}log_2(\frac{3}{3})] = 0$
europe:1+,2-	Entropy(S_{europe}) = $[-\frac{1}{3}log_2\left(\frac{1}{3}\right)-\frac{2}{3}log_2\left(\frac{2}{3}\right)]$= 0.9183
Gain(S, maker) $= Entropy(S) - \sum_{v\in\{asia,america,europe\}}(\lvert S_v\rvert/\lvert S\rvert)Entropy(S_v)$ $= 0.8631 - (\frac{1}{7}*0+\frac{3}{7}*0+\frac{3}{7}*0.9183) = 0.4696.$	

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 "



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