Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Công thức:

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

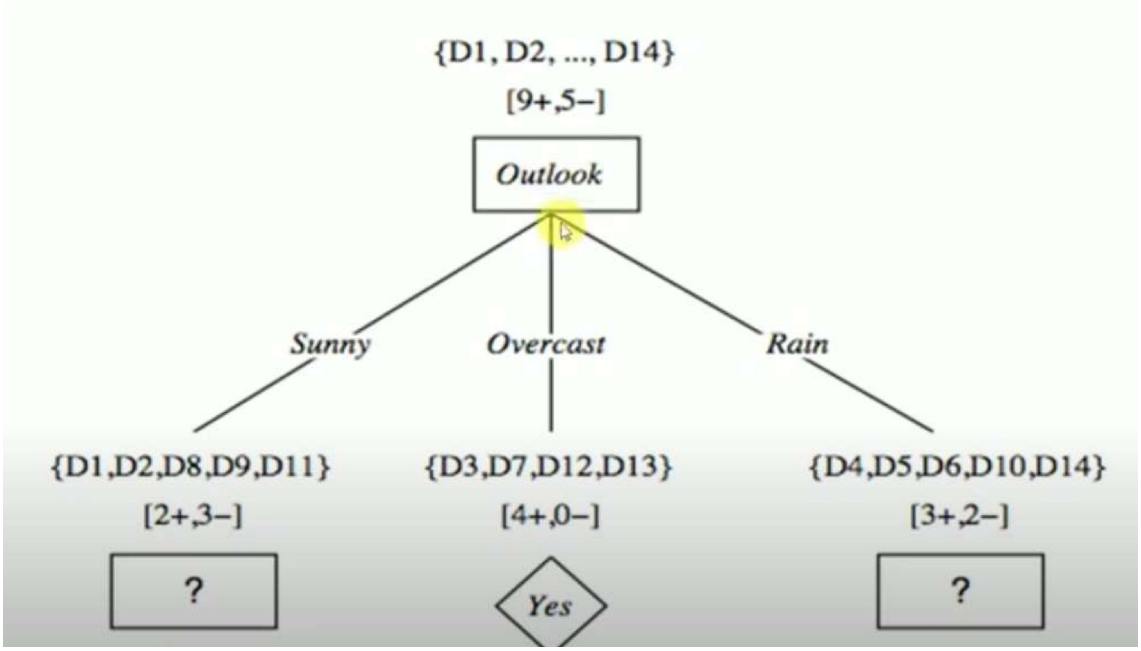
$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

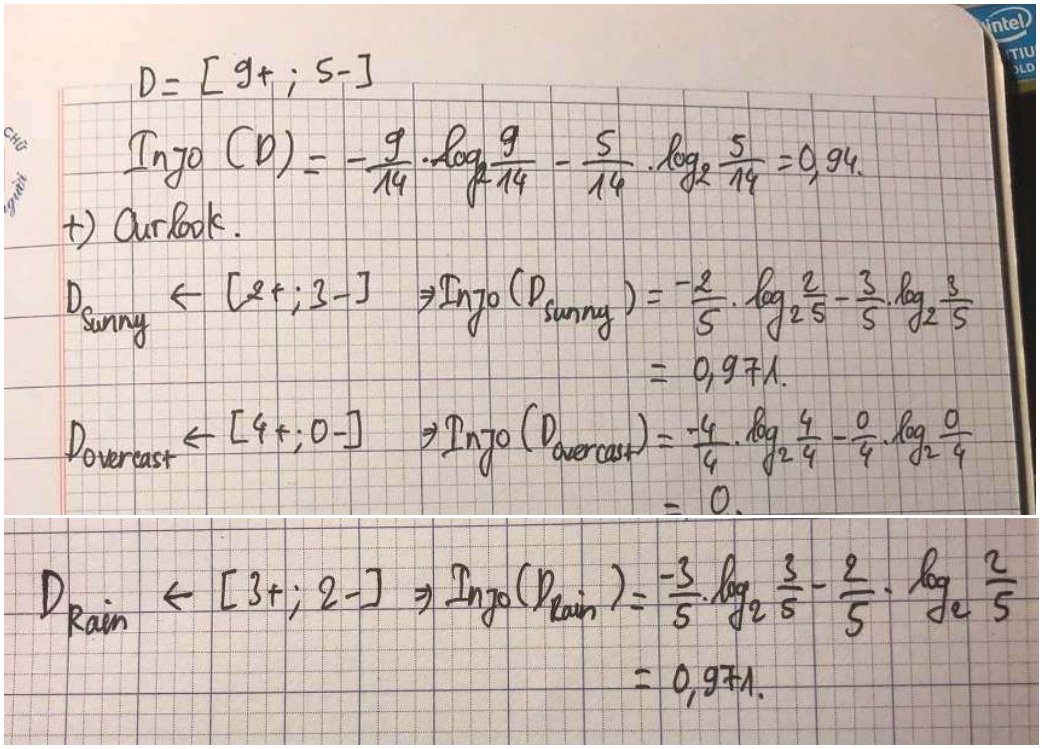
B1: Cần tìm Gain của 4 yếu tố: Outlook, Temp, Humidity, Wind để so sánh và xác định gốc trong decision tree.

- Thuộc tính “play tennis” là thuộc tính chính để tính Gain.
- “Play tennis” có 9 YES, 5 NO => D = [9+;5-].
- Tính Info (D) và tìm Gain của từng thuộc tính như trong ảnh. Thuộc tính có Gain cao nhất sẽ làm gốc trong decision tree.



Outlook làm gốc vì có chỉ số Gain cao nhất

- Các bước tính các Gain



$$Gain(D, Outlook) = Info(D) - \frac{5}{14}.Info(D_{sunny}) - \frac{4}{14}.Info(D_{overcast}) - \frac{5}{14}.Info(D_{rain}).$$
$$= 0,2464.$$

$= 0,2464.$

+) Temp.

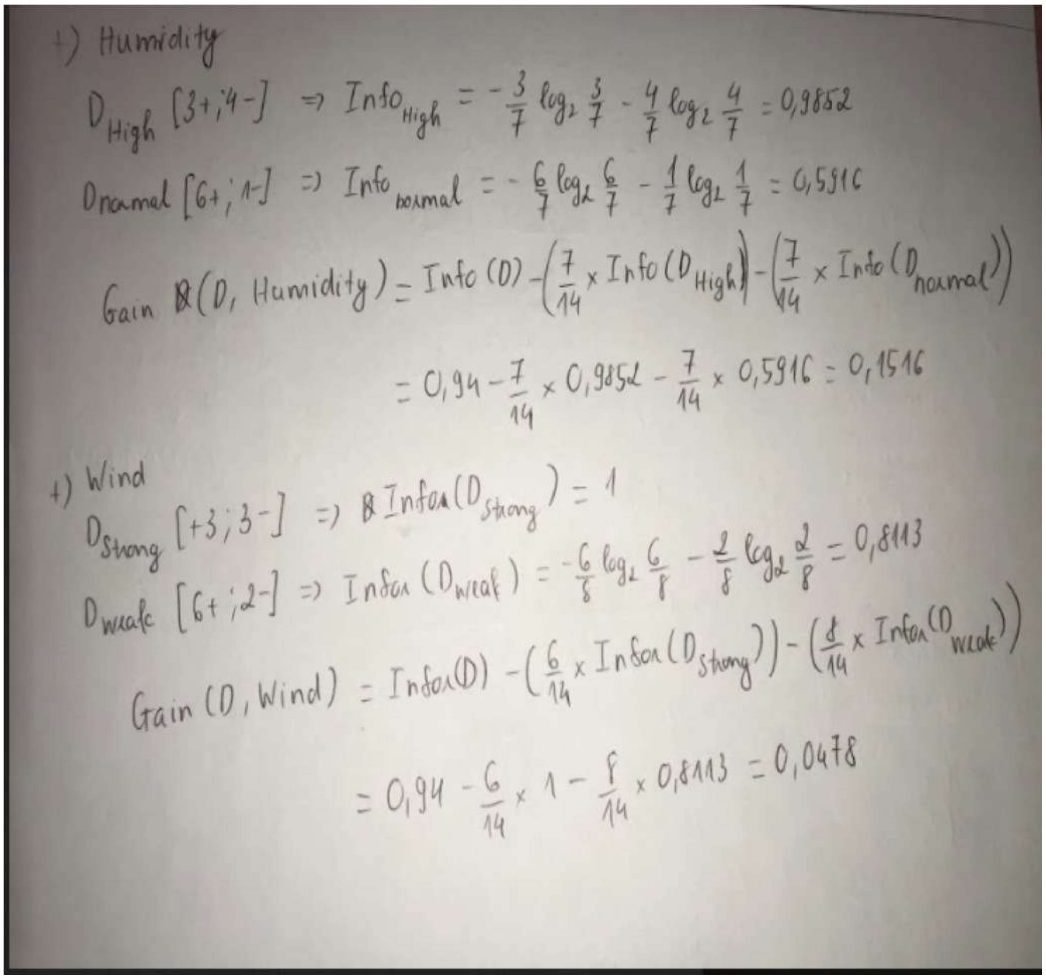
$D_{Hot} \leftarrow [9+; 5-] \Rightarrow Info(D_{Hot}) = \frac{-2}{4}.\log_2\frac{2}{4} - \frac{2}{4}.\log_2\frac{2}{4}$
$= 1.0.$

$D_{Mild} \leftarrow [4+, 2-] \Rightarrow Info(D_{Mild}) = \frac{-4}{6}.\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$
$= 0,9183.$

$D_{Cool} \leftarrow [3+, 1-] \Rightarrow Info(D_{Cool}) = \frac{-3}{4}.\log_2\frac{3}{4} - \frac{1}{4}.\log_2\frac{1}{4}$
$= 0,8113.$

$Gain(D, Temp) = Info(D) - \frac{4}{14}.Info(D_{Hot}) - \frac{6}{14}.Info(D_{Mild}) - \frac{4}{14}.Info(D_{Cool}) = 0,0289.$

HỒNG HÀ



B2: Tiếp tục tìm gốc của nhánh Sunny từ Outlook. Lọc những ngày Outlook = “Sunny”, ta được bảng như bên dưới:

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

Các bước làm tương tự như B1 để tìm gốc tiếp theo

+) Temp:

$D_{sunny} = [2+; 3-]$

$Info(D_{sunny}) = -\frac{2}{5}.\log_2\frac{2}{5} - \frac{3}{5}.\log_2\frac{3}{5} = 0,97.$

$D_{Hot} \leftarrow [0+; 2-] \Rightarrow Info(D_{Hot}) = 0,0,$

$D_{Mild} \leftarrow [1+; 1-] \Rightarrow Info(D_{Mild}) = 1,0.$

$D_{Cool} \leftarrow [1+; 0-] \Rightarrow Info(D_{Cool}) = 0,0.$

$Gain(D_{sunny}, Temp) = Info(D) - \frac{2}{5}.Info(D_{Hot})$

$- \frac{2}{5}.Info(D_{Mild}) - \frac{1}{5}.Info(D_{Cool})$

$= 0,57.$

+) Humidity:

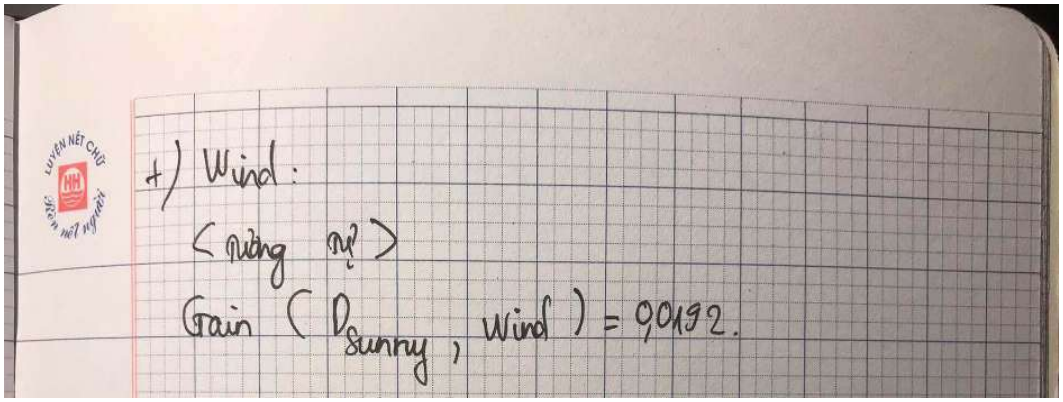
$D_{High} \leftarrow [0+; 3-] \Rightarrow Info(D_{High}) = 0,0.$

$D_{Normal} \leftarrow [2+; 0-] \Rightarrow Info(D_{Normal}) = 0,0.$

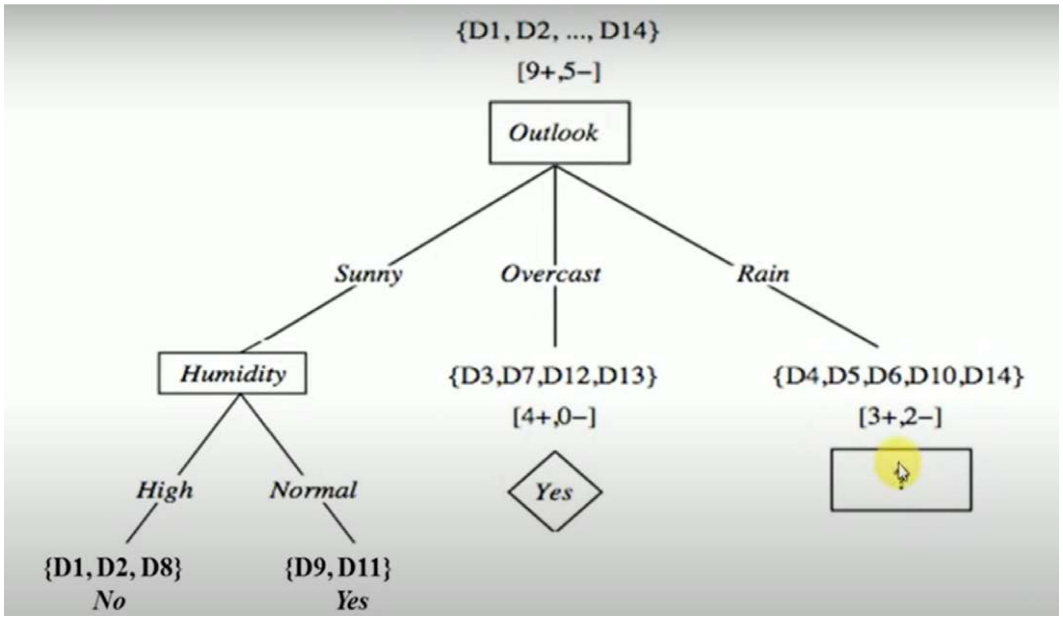
$Gain(D_{sunny}, Humidity) = Info(D) - \frac{3}{5}.Info(D_{High})$

$- \frac{2}{5}.Info(D_{Normal})$

$= 0,97.$



- Gain của Humidity là cao nhất nên Humidity tiếp tục làm gốc của nhánh tiếp theo.
- Dựa vào trong phần tính của Humidity (ở bước 2 này) , D high và D normal đều bằng không => Kết thúc (như ảnh dưới)



B3: Tiếp tục tương tự B2, tìm gốc của nhánh Rain từ Outlook.

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$$Gain(S_{Rain}, Wind) = 0.97$$

Ta có kết quả cuối cùng.