Decision Tree Using ID3 Algorithm

Windy?	Air Quality Good?	Hot?	Play Tennis?
No	No	No	No
Yes	No	Yes	Yes
Yes	Yes	No	Yes
Yes	Yes	Yes	No

Number of Yes = 2

Number of No's = 2.

Total Results = 4

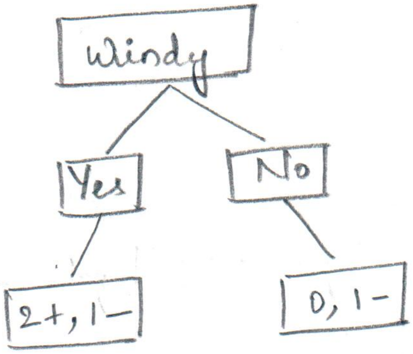
Class Label = Play Tennis

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

$$= -\left(\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right)$$

$$= 1$$

→ find the Information Gain for each attribute and then pick the attribute which provides the most information gain about the class label (Play Tennis)



$$Info_{Windy}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

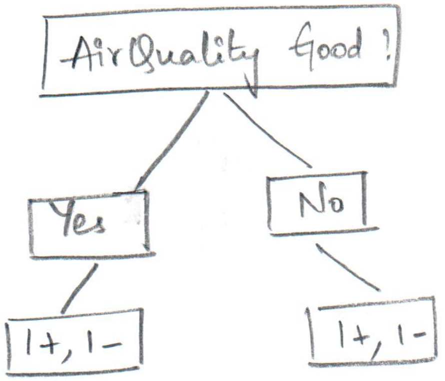
Yes = 3
No = 1

$$= \frac{3}{4}\left[-\left(\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right)\right)\right] +$$

$$\frac{1}{4}\left[-1 \log_2(1) + (0)\log_2(0)\right]$$

$$= 0.629$$

$$\boxed{Information\ Gain\ (D, Windy) = 1 - 0.629 = 0.371}$$



$$\text{Info}_{\text{Air Quality Good}} = \frac{2}{4}\left[-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] + \frac{2}{4}\left[-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right]$$

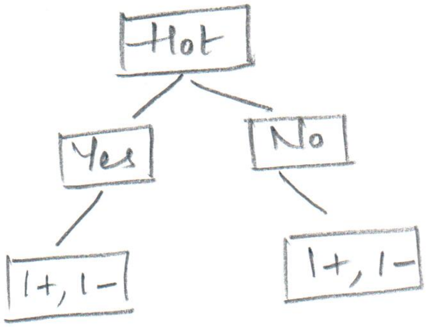
$$= 1$$

$$\boxed{\text{Gain}(D, \text{Air Quality Good}) = 1-1 = 0}$$

$$\text{Info}_{\text{Hot}} = \frac{2}{4}\left[-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right] + \frac{2}{4}\left[-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right]$$

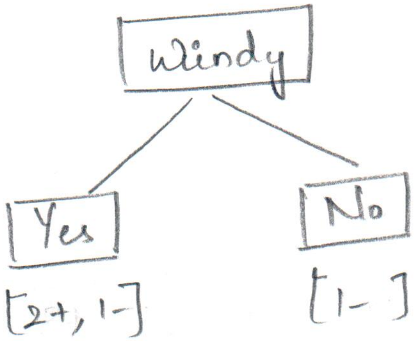
$$= 1$$

$$\boxed{\text{Gain}(D, \text{Hot}) = 1-1 = 0.}$$



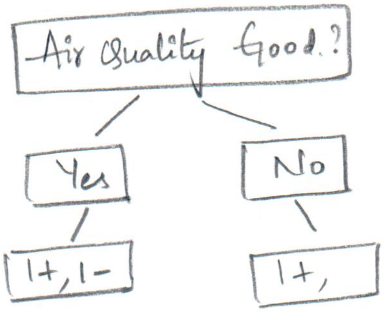
→ So, Here we choose Windy as root node because it provides the Information Gain.

→ Select the next attribute which provides highest Information Gain



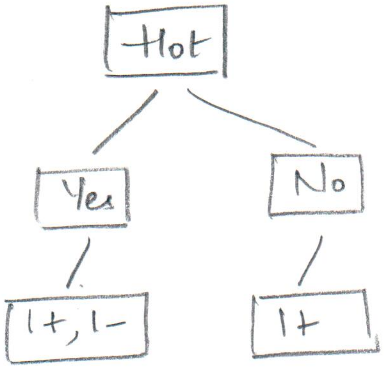
$$\text{Info}_{\text{Yes}}(D) = -\left[\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right)\right]$$

$$= 0.918$$



$$\text{Info}_{\text{Air Quality Good}} = \frac{2}{3}\left[-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] + \frac{1}{3}\left[-1\log_2(1)\right] = 0.666$$

$$\boxed{\text{Gain}(\text{Yes}, \text{Air Quality Good}) = 0.918 - 0.666 = 0.252}$$

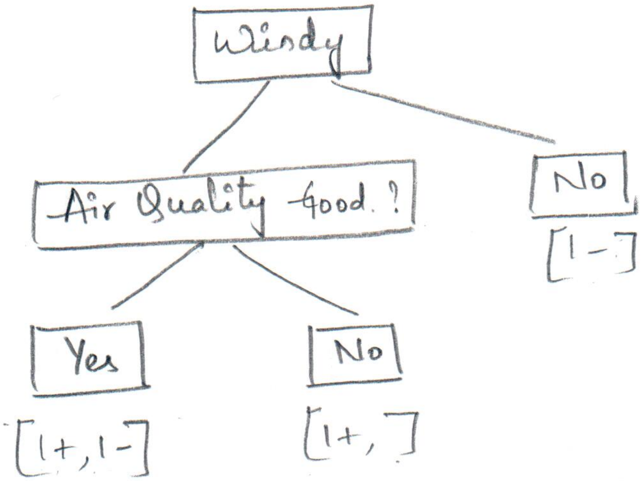


$$\text{Info}_{\text{Hot}} = \frac{2}{3}\left[-\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right] + \frac{1}{3}\left[-1.\log_2(1)\right]$$

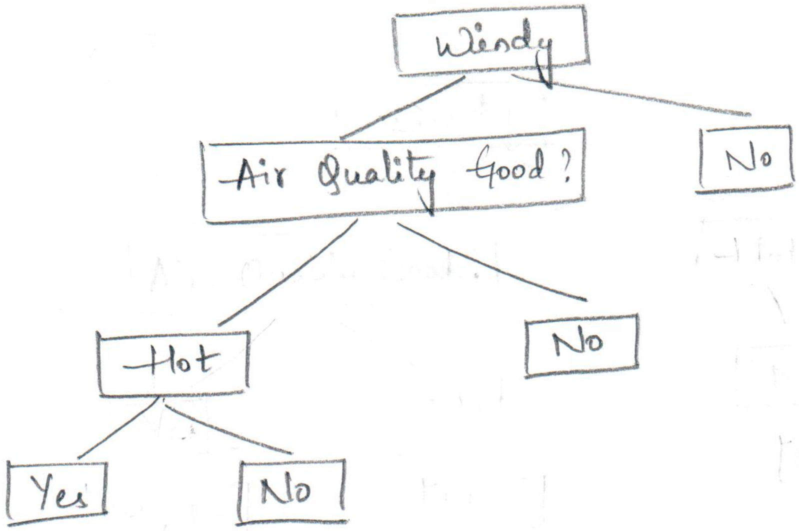
$$= 0.666$$

$$\boxed{\text{Gain}(\text{Yes}, \text{Hot}) = 0.918 - 0.666 = 0.252}$$

→ Here, we can choose either Air Quality or Hot as the attribute for Yes because both have the same Gain. I have chosen Air Quality.



→ We have only one attribute left, So, we will choose Hot as the attribute


Windy
Air Quality Good ?
No
Hot
No
Yes
No