DECISION TREE:-

⇒ Decision tree is a classification technique which is used to classify discrete value, target attributes.

⇒ The decision tree consist of root nodes, internal nodes, and leaf nodes.

⇒ The root node will be the best attribute. The best attribute can be derived based on the attribute selection measure like gain information or information gain based on Gini Index.

consider the dataset

Day	Weather	Temp	Windy	play
1	Rain	Mild	Weak	No
2	Normal	Hot	Weak	Yes
3	Wind	Mild	Strong	Yes
4	Normal	Cool	Weak	No
5	rain	Hot	Strong	No

construct or build the decision tree by finding the best attribute to decide whether to play or not

FORMULA: (c is the number of class labels)

$$\text{Entropy } E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$\text{Gain}(S_A) = \text{Entropy}(S) - \sum_{1}^{v} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

STEP-1:- For building a decision tree we need to find the entropy which is used to measure the impurity in the data attribute (Entropy E(S))

⇒ c is the number of class labels

We need to find the best attribute using information gain.

FORMULA FOR INFORMATION GAIN:- Gain (SA) given already the attribute with high information gain will be considered as the best attribute.

$$\text{Entropy}(S) = \sum_{i=1}^{2} -p_i \log_2 p_i$$

(2, 3)

2 yes 3 NO

$$= \left[-\frac{2}{5} \times \log_2 \frac{2}{5}\right] + \left[-\frac{3}{5} \times \log_2 3/5\right]$$

$$= -2/5 \times \log_2 \frac{2}{5} - \frac{3}{5} \times \log_2 \frac{3}{5} \times \log_2 \frac{3}{5}$$

$$= \frac{-2}{5} \times (-1.3219) - \frac{3}{5} \times (-0.736)$$

$$= 0.53 + 0.44 = \text{Entropy}(S) = 0.97$$

STEP-2:- Find Gain (information) for 3 attributes (weather, Temp, windy). The attribute with high information gain is best.

Gain(Weather)

Rain | Normal | Wind

2 | 2 | 1

0 2 | 1 1 | 1 0

"yes" "NO" | "yes" "NO" | "yes" "NO"

Entropy (weather = Rain) $\sum_{i=1}^{C} -p_i \log p_i$

(2)

0 2

"yes" "NO"

$= \frac{-0}{2} \times \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2}$

$= 0$.

Entropy (weather = Normal)

2

1 1

"yes" "NO"

$= \frac{-1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$

Entropy (weather = wind)

1

1 0

"yes" "NO"

$= -\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1}$

$= 0$.

Gain(weather) = Entropy (S) $- \sum_{v} \frac{|S_v|}{|S|}$ Entropy (S_v)

$= 0.97 - \frac{2}{5} \times 0 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0$

$= 0.97 - \frac{2}{5}$

Gain(Weather) = 0.57.

STEP-3:

Gain (Temp)

Gain(Temp)

Mild | Hot | Cool

2 | 2 | 1

"yes" "NO" | "yes" "NO" | "yes" "NO"

Entropy (Temp = Mild) $\sum_{i=1}^{c} -p_i \log p_i$

2

1 1

"yes" "No"

$= \frac{-1}{2} \log \frac{1}{2} - \frac{-1}{2} \log_2 \frac{1}{2}$

$= \frac{-1}{2} \times$

$0.5 + 0.5 = 1$

Entropy (Temp = Hot)

2

1 1

"yes" "No"

$= \frac{-1}{2} \log_2 \frac{1}{2} - \frac{-1}{2} \log_2 \frac{1}{2}$

$= 0.5 + 0.5 = 1$

Entropy (Temp = cool)

1

0 1

"yes" "No"

$= -\frac{0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1}$

$= 0$

Gain (Temp) = Entropy (S) $- \sum_{1}^{v} \frac{|S_v|}{|S|}$ Entropy (S_v)

$= 0.97 - \frac{2}{5} \times 1 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0$

Gain (Temp) = 0.17

STEP-4:-

Gain (Windy)

Weak Strong

3 2

1 2 1 1

"yes" "No" "yes" "No"

Entropy (Of Windy = Weak) $= \sum_{i=1}^{c} -p_i \log p_i$

3

1 2

"yes" "No"

$= \frac{-1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}$

$= 0.52 + 0.38 = (0.154) + 0.38$

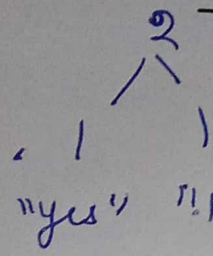
$= 0.91$

Entropy

Gain (windy) = $-\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$

= strong

= 0.5 + 0.5 = 1



Gain (windy) = $0.97 - \frac{3}{5} \times 0.91 - \frac{2}{5} \times 1$

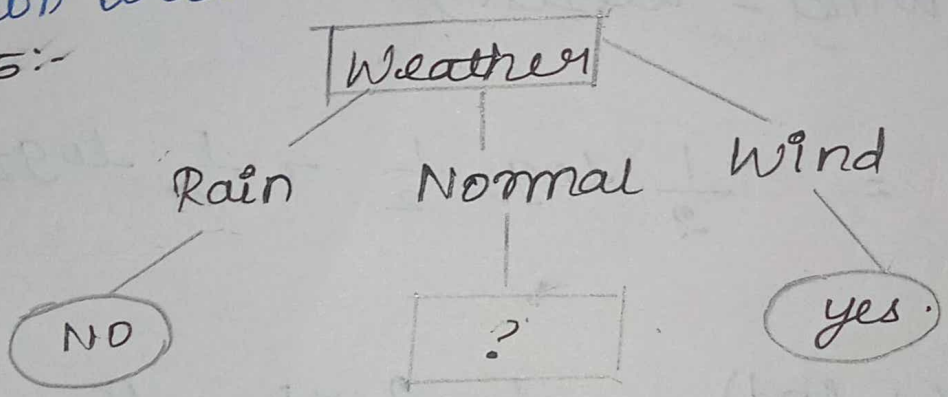
$0.97 - 0.546 - 0.4 = 0.024$.

Gain Info (Weather) = 0.57
Gain Info (Temperature) = 0.17
Gain Info (Wind) = 0.024

"Weather" ⇒ The highest gain information value is "0.57" which represents the attribute Weather.

The Weather attribute is the root of the decision tree.

STEP-5:-



(yes or NO we for Norm
(cannot conclude

(To find the next attribute)

New dataset S:

Temperature	Wind	play
Hot	Weak	Yes
cool	Weak	No

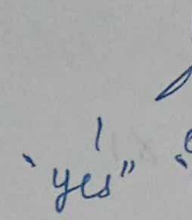
Entropy(S) = $\frac{-1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$

2 → 1 "yes", 1 "No"

= 1

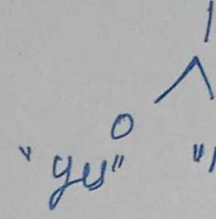
Gain Info (Temperature) = {Hot, Cool}

Entropy (Temp = "Hot")



$= \frac{-1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0$

Entropy (Temp = "cool")



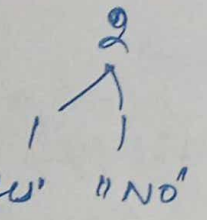
$= \frac{-0}{01} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} = 0$

Gain Info (Temperature) = Entropy (S) $- \sum_{1}^{v} \frac{|S_v|}{|S|}$ Entropy (S)

$= 1 - \frac{1}{2} \times 0 - \frac{1}{2} \times 0.$

GainInfo (wind) = {weak}.

Entropy (wind = "weak")



$= \frac{-1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1.$

Gain Info (wind) = $1 - \frac{2}{2} \times 1 = 0.$

Now,

GainInfo (Temp) = 1

GainInfo (wind) = 0

Here, Gain Info of 'temp' is higher, so we take temperature as the next attribute.

Decision tree: