Mon Tue Wed Thu Fri Sat

تاریخ:

Topic 07:- Decision Tree

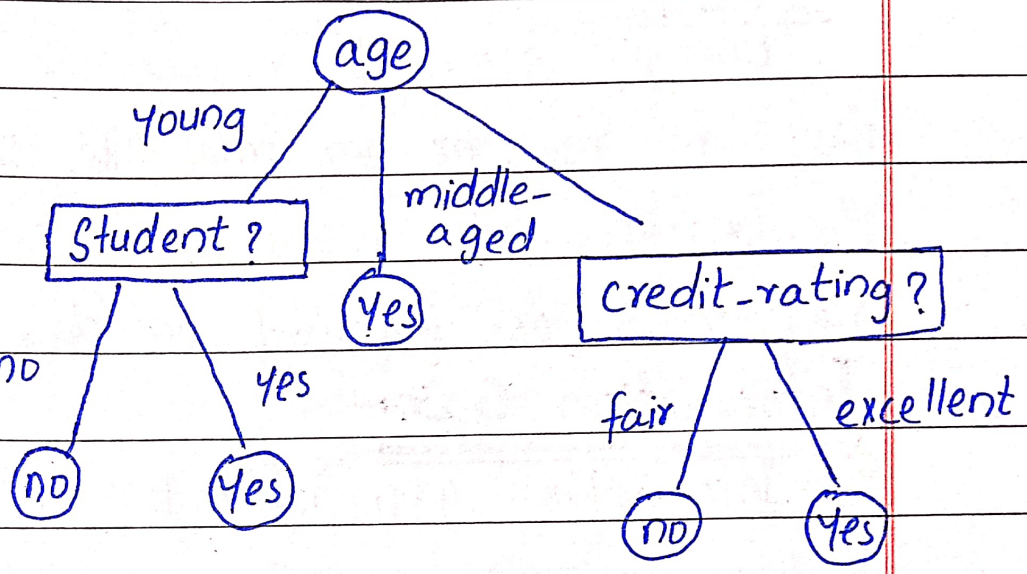
Decision Tree:-

A decision tree is a structure that includes a root node, branches and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

Example:-

The following decision tree is for the concept buy-computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node

represents a class.



Decision Tree Induction:-

It is a supervised learning method used in data mining for classification and regression methods. It helps us in decision making purposes. It creates classification or regression models. It separates a data set into smaller subsets, and at the same time, the decision tree developed steadily. The final tree is with decision nodes and leaf node.

Mon Tue Wed Thu Fri Sat

Key factors:-

Entropy:-

Entropy refers to a common way to measure impurity. In the decision tree, it measures the randomness or impurity in data sets.

Information Gain:-

Information gain refers to the decline in entropy after the data set is split. It is also called Entropy Reduction. Building a decision tree is all about discovering attributes that return the highest data gain.

Why are decision tree useful?

It enables us to analyze the possible consequences of a decision throughly.

It provides us framework to measure the values of outcomes and the probability of accomplishing

them.

Example:-

Day	Outlook	Temp	Humidity	Wind	Decision
1	sunny	Hot	High	Weak	No
2	sunny	Hot	High	Strong	No
3	overcast	Hot	High	weak	Yes
4	Rain	Mild	High	weak	Yes
5	Rain	cool	Normal	weak	Yes
6	Rain	cool	Normal	strong	No
7	overcast	cool	Normal	strong	Yes
8	sunny	Mild	High	weak	No
9	sunny	cool	Normal	weak	Yes
10	Rain	Mild	Normal	weak	Yes
11	Sunny	Mild	Normal	strong	Yes
12	Overcast	Mild	High	strong	Yes
13	Overcast	Hot	Normal	weak	Yes
14	Rain	Mild	High	strong	No

Mon Tue Wed Thu Fri Sat

تاریخ:

Step 1:-

Find information gain of target attributes.

$$I.G = -\frac{P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

$$\text{Entropy}(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= 0.410 + 0.530$$

$$= 0.94$$

Step 2:-

Find entropy of remaining attributes one by one.

Entropy (S, outlook):-

	Yes	No	Total
sunny	3	2	5
overcast	4	0	4
rainy	2	3	5

Mon Tue Wed Thu Fri Sat

تاریخ:

$$E(A) = \sum_{i=1}^{r} \frac{P_i + N_i}{P+N}, \frac{I(P_i N_i)}{I.G}$$

$E(S, \text{outlook})$ $P=2$ $N=3$ sunny $= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.529 + 0.442$

$= 0.971$

$E(S, \text{outlook})$ $P=4$ $N=0$ overcast $= -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$

$= 0$

$E(S, \text{outlook})$ $P=3$ $N=2$ rain $= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.442 + 0.529$

$= 0.971$

$E(S, \text{outlook}) = \frac{5}{14}(E(3,2)) + \frac{4}{14}(E(4,0)) + \frac{5}{14}(E(3,2))$.

$= \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$

$= 0.6936$

Step 3:- Calculate Gain

Mon Tue Wed Thu Fri Sat

Gain (S, Outlook)

Gain = I.G - E(A)

= 0.94 - 0.6936

= 0.2464.

Now, we will repeate step 2 & 3 for other attributes.

Entropy (S, temperature):-

	Yes	No	Total
Hot	2	2	4
Mild	4	2	6
Cool	3	1	4

(2,2)

$$E(S, \text{temperature})_{\text{hot}} = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 0.5 + 0.5$$

(4,2) $$= 1$$

$$E(S, \text{temperature})_{\text{mild}} = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$= 0.390 + 0.528$$

$$= 0.918$$

(3,1) $$E(S, \text{temperature})_{\text{cool}} = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$= 0.311 + 0.5$

$= 0.811$

$E(S, \text{temperature}) = \frac{4}{14}(E(2,2)) + \frac{6}{14}E(4,2) + \frac{4}{14}E(3,1))$

$= \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811)$

$= 0.911$

Gain(S, Temprature):-

$= 0.94 - 0.911$

$= 0.029$

Entropy (S, Humidity):-

	Yes	No	Total
High	3	4	7
Normal	6	1	7

(3,4)

$E(S, \text{Humidity})_{\text{high}} = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$

$= 0..524 + 0.4614$

$= 0.985$

Mon Tue Wed Thu Fri Sat

تاریخ:

(6,1)

$$E(S, \text{Humidity})_{\text{normal}} = \frac{-6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

$$= 0.1906 + 0.401$$

$$= 0.592.$$

$$E(S, \text{humidity}) = \frac{7}{14}(E(3,4)) + \frac{7}{14}(E(6,1))$$

$$= \frac{7}{14}(0.9852) + \frac{7}{14}(0.592)$$

$$= 0.788$$

Gain (S, Humidity) :-

$$= 0.94 - 0.788$$

$$= 0.152$$

Entropy (S, Wind) :-

	Yes	No	Total
weak	6	2	8
strong	3	3	6

(6,8)

$$E(S, \text{wind})_{\text{weak}} = \frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$$

$$= 0.311 + 0.5$$

$$= 0.811$$

Mon Tue Wed Thu Fri Sat

$E(S, wind) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\frac{3}{6}$
(3,3) Strong

$= 0.5 + 0.5$

$= 1$

$E(S, wind) = \frac{8}{14}(E(6,2)) + \frac{6}{14}(E(3,3))$

$= \frac{8}{14}(0.811) + \frac{6}{14}(1)$

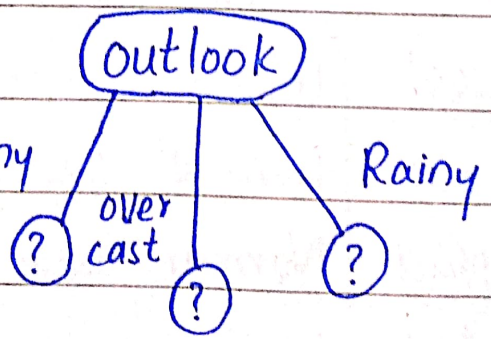
$= 0.892$

Gain (S, wind):-

$= 0.94 - 0.892$

$= 0.048$

From gain of all four attributes Gain (S, outlook) has maximum value, so **Outlook** will be the **root node** of this decision tree.



Mon Tue Wed Thu Fri Sat

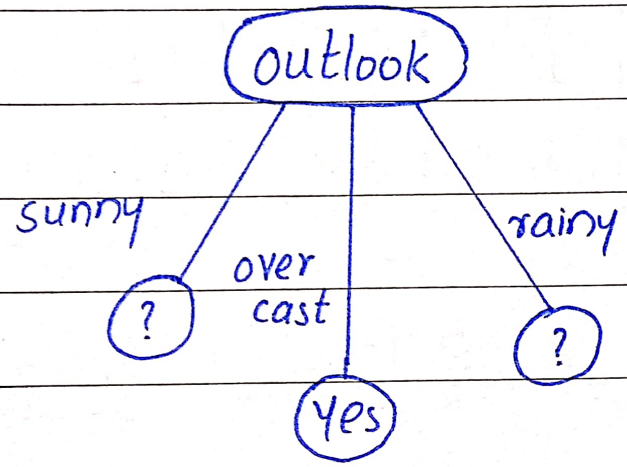
As outlook have three values, so we will divide given table into three tables according to each value.

Outlook	Temp	Humidity	Wind	Decision
sunny	Hot	High	weak	No
sunny	Hot	High	strong	No
sunny	Mild	High	weak	No
sunny	Cool	Normal	weak	Yes
sunny	Mild	Normal	strong	Yes

Outlook	Temp	Humidity	Wind	Decision
overcast	Hot	High	weak	Yes
overcast	Cool	Normal	strong	Yes
overcast	Mild	High	strong	Yes
overcast	Hot	Normal	weak	Yes

Outlook	Temp	Humidity	Wind	Decision
Rain	Mild	High	weak	Yes
Rain	Cool	Normal	weak	Yes
Rain	Cool	Normal	strong	No
Rain	Mild	Normal	weak	Yes
Rain	Mild	High	strong	No

Since overcast contains only examples of class 'Yes' we can set it as Yes. That means if outlook is overcast match will be played.



Now, we will solve 1st table:

Sunny :-

1st Step :-

$$E(Sunny) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= 0.442 + 0.529$$

$$= 0.971$$

Step 2 & 3 for all attributes:-

E (Sunny, Temperature):-

	Yes	No	Total
Hot	0	2	2
Mild	1	1	2
Cool	1	0	1

$$E(S,T)_{hot} = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$= 0$$

$$E(S,T)_{Mild} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= 0.5 + 0.5$$

$$= 1$$

$$E(S,T)_{cool} = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)$$

$$= 0$$

E (Sunny, Temperature) =

$$\frac{2}{5}(E(0,2)) + \frac{2}{5}(E(1,1)) + \frac{1}{5}(E(1,0))$$

$$= \frac{2}{5}(0) + \frac{2}{5}(1) + \frac{1}{5}(0) = 0.4$$

Mon Tue Wed Thu Fri Sat

تاریخ:

Gain (Sunny, Temperature):-

$= 0.971 - 0.4$

$= 0.571$

E (Sunny, Humidity):-

	Yes	No	Total
High	0	3	3
Normal	2	0	2

$$E(S,H)_{high} = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right)$$

$$= 0$$

$$E(S,H)_{normal} = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{2}{2}\right)$$

$$E(Sunny, humidity) = \frac{3}{5}(E(0,3)) + \frac{2}{5}(E(2,0))$$

$$= \frac{3}{5}(0) + \frac{2}{5}(0) = 0$$

Gain (Sunny, Humidity):-

$= 0.971 - 0$

$= 0.971$

Mon Tue Wed Thu Fri Sat

تاریخ: _____

E(Sunny, Wind):-

	Yes	No	Total
strong	1	1	2
Weak	1	2	3

$$E(S,W)_{weak} = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$= 0.528 + 0.390$$

$$= 0.918$$

$$E(S,W)_{strong} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= 0.5 + 0.5$$

$$= 1$$

$$E(\text{Sunny, Wind}) = \frac{2}{5}(E(1,1)) + \frac{3}{5}(E(1,2))$$

$$= \frac{2}{5}(1) + \frac{3}{5}(0.918)$$

$$= 0.9508$$

Gain (Sunny, Wind):-

$$= 0.971 - 0.9508$$

$$= 0.020$$

Here, Gain (Sunny, Humidity) is the largest value. So **humidity** is the node that comes under sunny.

Now, we will form the table for humidity if outlook is sunny.

	Yes	No	Total
High	0	3	3
Normal	2	0	2

If outlook is sunny and humidity is high, then decision will be No, and if humidity is Normal, then decision will be Yes.

Now, we will solve 3rd table.

Mon Tue Wed Thu Fri Sat

Rain :-

1st Step :-

$$E(\text{Rain}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= 0.971$$

Step 2 & 3 for all attributes :-

E (Rain, Temperature) :-

	Yes	No	Total
Hot	0	0	0
Mild	2	1	3
Cool	1	1	2

$$E(R,T)_{\text{mild}} = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$= 0.390 + 0.528$$

$$= 0.918$$

$$E(R,T)_{\text{cool}} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= 0.5 + 0.5$$

$$= 1$$

Mon Tue Wed Thu Fri Sat

تاریخ:

$$E(\text{Rain, Temperature}) = 0 + \frac{3}{5}(E(2,1)) + \frac{2}{5}(E(1,1))$$

$$= 0 + \frac{3}{5}(0.918) + \frac{2}{5}(1)$$

$$= 0.9508$$

Gain (Rain, Temperature):-

$$= 0.971 - 0.9508$$

$$= 0.020$$

E (Rain, Humidity):-

	Yes	No	Total
High	1	1	2
Normal	2	1	3

$$E(R,H)_{high} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= 0.5 + 0.5 = 1$$

$$E(R,H)_{normal} = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$= 0.390 + 0.528 = 0.918$$

Mon Tue Wed Thu Fri Sat

$E(\text{Rain, Humidity}) = \frac{2}{5}(E(1,1)) + \frac{3}{5}(E(2,1))$

$= \frac{2}{5}(1) + \frac{3}{5}(0.918)$

$= 0.9508$

Gain (Rain, Humidity):-

$= 0.971 - 0.9508$

$= 0.020$

E (Rain, Wind):-

	Yes	No	Total
Strong	0	2	2
Weak	3	0	3

E(Rain, Wind) = 0

Gain (Rain, Wind):-

$= 0.971 - 0$

$= 0.971$

Here, Gain (Rain, Wind) has the maximum value.

Mon Tue Wed Thu Fri Sat

تاریخ: _____

So, It is clear that if outlook is rain then the next node will be **Wind.**

If wind is strong, then decision will be No. And if wind is weak, then decision will be Yes.

Decision Tree:-