Date _____ Day _____

"DECISION TREE"

- Information Gain measures how well a given attribute / feature seperates the training examples acc to their target classification.

- ID3 algo uses information gain (entropy) to measure to select among the candidate attributes at each step while making the tree.

- "Best Attribute" is the one with lowest entropy or highest IG.

- Entropy is the measure of "randomness" or "disorder"
- Entropy is 0 if all members of a collection S belongs to same class. (like all values are Yes) because there is no randomness

- "The diff in the entropy before & after the split is Information Gain (IG)"

Date _____ Day _____

$$E(S) = -P(Yes)\log_2(P(Yes)) - P(No)\log_2(P(No))$$
$$= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$
$$= 0.940$$

Now calculating "E" for each attribute

$\boxed{E(\text{Outlook}):}$

Play Tennis = Yes

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

total no. of Sunny

$$= 0.9709$$

$$E(\text{Outlook} = \text{Overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$$
$$= 0$$

$$E(\text{Outlook} = \text{Rain}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$
$$= 0.9709$$

$$IG = 0.940 - \left[0.9709 \times \frac{5}{14} + 0.9709 \times \frac{5}{14}\right]$$
$$IG = 0.2465$$

Date _____ Day _____

$\boxed{E(Temp):}$

$$E(Temp = Hot) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$= \cancel{0.8112}\ 1$$

$$E(Temp = Mild) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$$

$$= 0.9183$$

$$E(Temp = Cool) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 0.8112$$

$$IG = 0.940 - \left[4/14 \times \overset{1}{\cancel{0.8112}} + 6/14 \times \overset{0.9183}{0.\cancel{8112}} + 4/14 \times 0.8112\right]$$

$$IG = 0.0289$$

Date _____ Day _____

E(Humidity):

$$E(\text{Humidity} = \text{High}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$

$$= 0.985$$

$$E(\text{Hum} = \text{Normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

$$= 0.592$$

$$IG = 0.940 - [7/14 * 0.985 + 7/14 * 0.592]$$

$$IG = 0.1515$$

E(Wind):

$$E(\text{Wind} = \text{Weak}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$$= 0.8113$$

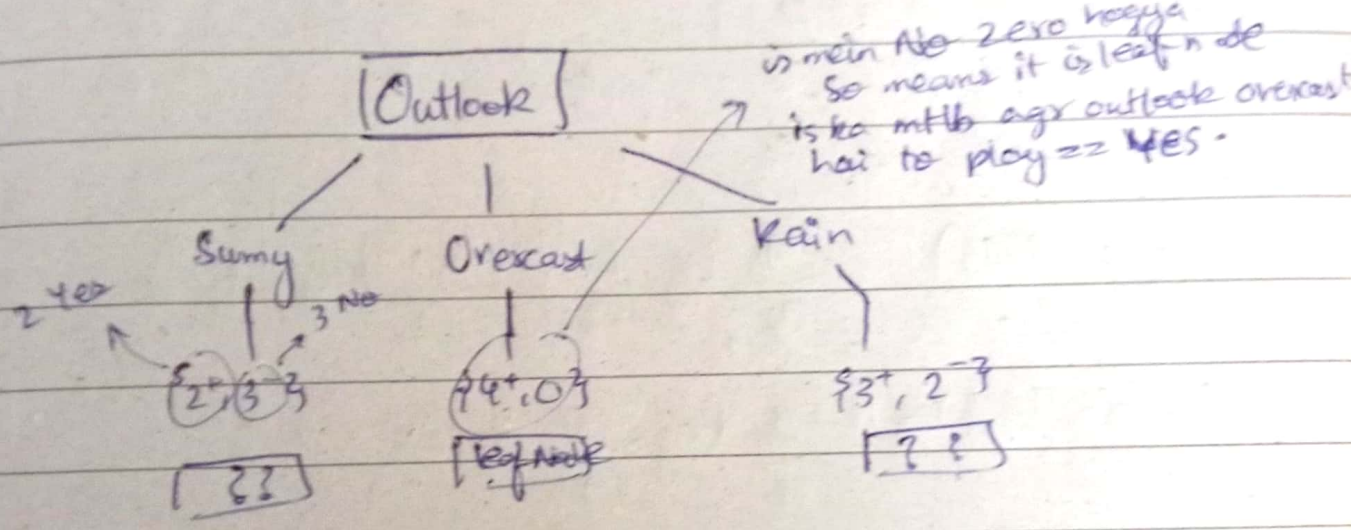
$$E(\text{Wind} = \text{Strong}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

$$= 1$$

Date _____ Day _____

$IG = 0.940 - [8/14 \times 0.8113 + 6/14 \times 1]$
$= 0.047$

So highest IG is of Outlook i.e 0.246
Root Node = Outlook



Day	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sun	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Norm	Weak	Yes
11	Sunny	Mild	Norm	Strong	Yes

ab is table k liye
krna hai.

Date _____ Day _____

total yes in that table

total samples in sub table

$$E(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$E(S) = 0.9710$$

E(temperature):

$$E(temp = Hot) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$= 0$$

$$E(temp = Mild) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= 1$$

$$E(temp = Cool) = -\frac{1}{1}\log_2\frac{1}{1} - 0 = 0$$

$$IG = 0.9710 - [0 + 2/5 \times 1 + 0]$$

$$\boxed{IG = 0.5710}$$

E(Humidity):

$$E(Hu = High) = \frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3}$$

$$= 0$$

$$E(Hu = Normal) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log\frac{0}{2}$$

$$= 0$$

Date _____ Day _____

$IG = 0.971 - [0+0]$

$\boxed{= 0.971}$

E(Wind) :

$E(Wind = Strong) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

$= 1$

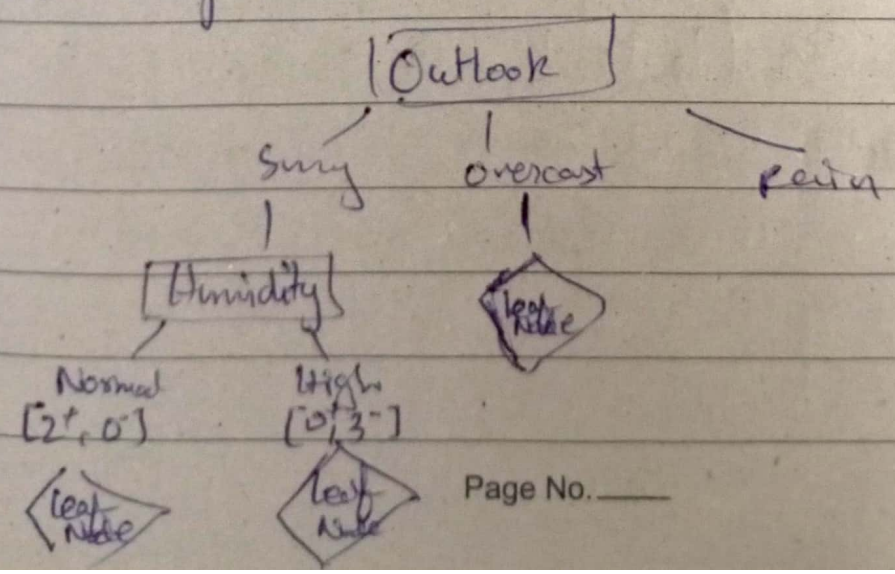
$E(W = Weak) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.9183$

$IG = 0.971 - [2/5 \times 1 + 3/5 \times 0.9183]$

$\boxed{IG = 0.2}$

For Sunny Next Node will be Humidity



Page No. _____

Date Day

Now making table for Rain.

Day	Outlook	Temp	Humidity	Wind	Play
4	Rain	Mild	High	Weak	Yes
5	"	Cool	N	Weak	Yes
6	"	Cool	N	Strong	No
10	"	Mild	N	Weak	Yes
14	"	Mild	High	Strong	No

$E(S) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}(\log_2(\frac{2}{5})$

$= 0.9710$

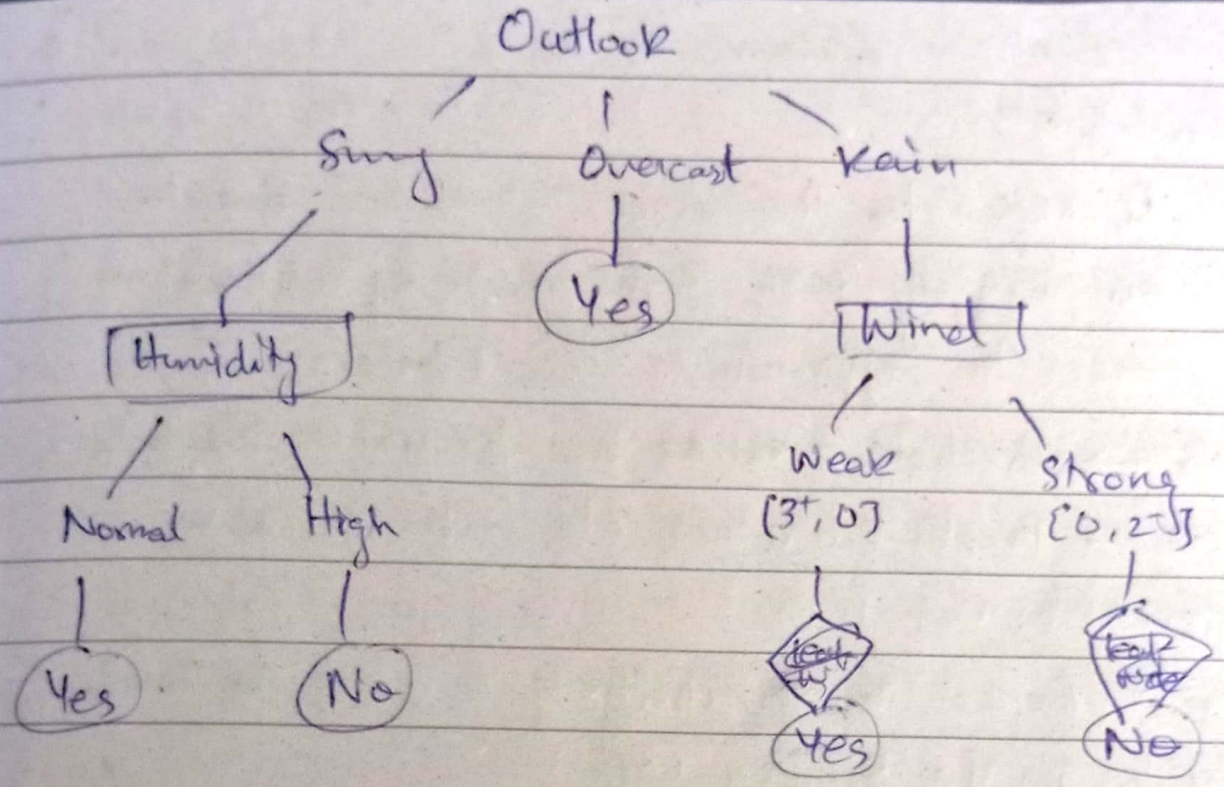
$IG(Wind) = 0.971 - \frac{3}{5}(\frac{3}{5}\log_2\frac{3}{5} - 0) - \frac{2}{5}(0 - \frac{2}{5}\log\frac{2}{5}$

$= 0.971$ ✓

$IG(Temp) = 0.971 - \frac{3}{5}(\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2(\frac{1}{3})) - \frac{2}{5}(\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2})$

$= 0.007$

Date _____ Day _____



→ CART → Classification & Regression Trees

→ ID3 performs Hill climbing search through hypothesis space

→ DT has problem of overfitting

↳ It grows deeply enough to perfectly classify all training example. This leads to problem when there is noise in data.

Cat classification example

	Ear shape (x_1)	Face shape (x_2)	Whiskers (x_3)	Cat
1	Pointy	Round	Present	1
2	Floppy	Not round	Present	1
3	Floppy	Round	Absent	0
4	Pointy	Not round	Present	0
5	Pointy	Round	Present	1
6	Pointy	Round	Absent	1
7	Floppy	Not round	Absent	0
8	Pointy	Round	Absent	1
9	Floppy	Round	Absent	0
10	Floppy	Round	Absent	0

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

ID	Fever	Cough	Breathing issues	Infected
1	NO	NO	NO	NO
2	YES	YES	YES	YES
3	YES	YES	NO	NO
4	YES	NO	YES	YES
5	YES	YES	YES	YES
6	NO	YES	NO	NO
7	YES	NO	YES	YES
8	YES	NO	YES	YES
9	NO	YES	YES	YES
10	YES	YES	NO	YES
11	NO	YES	NO	NO
12	NO	YES	YES	YES
13	NO	YES	YES	NO
14	YES	YES	NO	NO

→ classification algo

Decision Tree

Date:

Example: Infected patient diagnosis.

→ Every node is a feature/attribute.

→ ~~Phele node k liye esa attri~~

(1) Calculate entropy of whole data set.

measure of impurity

$E(S) = H(S) = ?$

$$E(S) = -P(yes)\log_2(P(yes)) - P(No)\log_2(P(No))$$

$$= \frac{8}{14}\log_2\left(\frac{8}{14}\right) - \frac{6}{14}\log_2\left(\frac{6}{14}\right)$$

$$E(S) = 0.985$$

All this work is done to select root node.

(2) Calculate "E" for each attribute

E(fever) → Fever k col mein total 2 classification hai

Yes or No done par check kara hai.

For Yes: → Jitne logo ko fever hai unmein se infected kitne hai wo E(S) k formula mein Yes mein aayenge or jitne ko infection nhi wo No mein.

$$P(Fever=Yes) = \frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.81$$

Where Fever = No & Infected = Yes

Fever = No & Infected = No

can No mein kitne infected = Yes

Date:

$$P(\text{Fever}=\text{No}) = -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)$$

total No in fever

total No in fever

$= 0.91$

$$\text{Information Gain (Fever)} = H(S) - \left[H\left(\text{Fever}=\text{Yes}\right) \times \frac{\text{Total Fever}=\text{Yes}}{\text{Total Samples}} + H\left(\text{Fever}=\text{No}\right) \times \frac{\text{Total Fever}=\text{No}}{\text{Total samp}}\right]$$

$$= 0.985 - \left[0.81 \times \frac{8}{14} - 0.91 \times \frac{6}{14}\right]$$

$$\boxed{\text{I.G (Fever)} = 0.13}$$

$H(\text{Cough})$:

dono same hai to 1 hoti hai entropy

$$(\text{Cough}=\text{Yes}) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$$

$= 1$

$$H(\text{Cough}=\text{No}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$= 0.811$

Date:

$$I.G(Cough) = 0.985 - \left[1 \times \frac{10}{14} + 0.811 \times \frac{4}{14}\right] = 0.04$$

$$IG(Cough) = 0.04$$

H (Breath Issue):

$$H(B=Yes) = \frac{-7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) = 0.54$$

$$=$$

$$H(B=No) = -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{5}{6}\log_2\left(\frac{5}{6}\right) = 0.65$$

$$I.G(BI) = 0.985 - \left[0.54 \times \frac{8}{14} + 0.65 \times \frac{6}{14}\right] = 0.4$$

$$IG(Breath\ Iss) = 0.4$$

▶ Jiski sabse zyada information gain hoga wo tree ka root node hoga.
So it will be Breathing Issue

Cat classification

Date:

$$E(S) = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$$

$$\boxed{E(S) = 1}$$

E (Ear shape):

$$E(ES = \text{Pointy}) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)$$

$$= 0.722$$

$$E(ES = \text{Floppy}) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)$$

$$= 0.722$$

$$IG(ES) = 1 - \left[0.722 \times \frac{5}{10} + 0.722 \times \frac{5}{10}\right]$$

$$= 0.28$$

E (Face Shape):

$$E(FS = \text{Round}) = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right)$$

$$= 0.985$$

$$E(ES = \text{Not Round}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

Date:

$= 0.918$

$IG(PS) = 1 - \left[0.905 \times \frac{7}{10} + 0.918 \times \frac{3}{10}\right]$

$IG(PS) = 0.351$

E(Whiskers):

$E(W = Present) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= 0.811$

$E(W = Absent) = -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)$

$= 0.918$

$IG(W) = 0.1248$

→ To limit depth we do pruning.
↳ pre pruning (train kone se phele cutoff kerenge)
↳ post pruning (after training whole model we cut off features)

Date:

→ generalized model or stable model we hote hai k small change bhi krein to koi effect na ho.

K-Means Clustering

x_1	x_2
1	1
1.5	2
3	4
5	7
3.5	5
4.5	5

$O(n \cdot k \cdot t)$

n → no. of data points
k → no. of centroids
t → no. of iterations

hamara point jisk close hoga wo hoga (1.5, 2) ek se close hai bcz dist is less

→ each point to a cluster

- generalize model
  ↳ jismein bias variance dono kam ho.

→ jitne features utni tree ki depth.

→ jitni depth utni overfitting & takes more computational power.

Date:

→ Decision tree has overfit limitation

→ Bias Variance

↳ Training data error

↳ Testing data error

→ Decision tree has high variance

↳ Underfit mein Low

↳ overfit