Dataset: Weather to Play Tennis

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

Scanned with CamScanner

ID3 Algorithm - Numerical

Dataset : Weather

Decision Target: Play Tennis or Not

Probability Table

	⊕	⊖		⊕	⊖		⊕	⊖		⊕	⊖		⊕	⊖
Sunny (5)	2	3	Hot (4)	2	2	High (7)	3	4	Weak (8)	6	2			
Overcast (4)	4	0	Mild (6)	4	2	Normal (7)	6	1	Strong (6)	3	3	Total (14)	9	
Rain (5)	3	2	Cool (4)	3	1									
	OUTLOOK			TEMP			HUMIDITY			WIND				

1) $\text{Entropy}_{(total)} = -\left[\frac{9}{14} * \log_2\left(\frac{9}{14}\right)\right] - \left[\frac{5}{14} * \log_2\left(\frac{5}{14}\right)\right] = \boxed{0.94029}$

Calculate the Information gain for four features
- Outlook
- Temp
- Humidity
- Wind

2) $\boxed{\text{Gain}(\underbrace{S_{total}}_{\text{All Samples}}, \text{Outlook}) = ?}$

Outlook Values → Sunny, Overcast, Rain

$$= \text{Entropy}_{(total)} - \sum \left| \frac{N_{value}}{N_{total}} * \text{Entropy}_{value} \right|$$

2a) $\text{Entropy}_{sunny} = -\frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{3}{5} * \log_2\left(\frac{3}{5}\right) = \boxed{0.97095}$

2b) $\text{Entropy}_{overcast} = -\frac{4}{4} * \log_2\left(\frac{4}{4}\right) - \left(\frac{0}{4}\right) * \log_2\left(\frac{0}{4}\right) = \boxed{0}$

2c) $\text{Entropy}_{rain} = -\frac{3}{5} * \log_2\left(\frac{3}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right) = \boxed{0.97095}$

Scanned with CamScanner

$$Gain(S_{total}, Outlook) = 0.94029 - \left[\frac{5}{14} * 0.97095 + \frac{4}{14} * 0 + \frac{5}{14} * 0.97095\right]$$

$$= \boxed{0.24675}$$

3) $\boxed{Gain(S_{total}, Temp) = ?}$

Temp Values → Hot, Mild, Cool

3a) $E_{hot} = -\left[\frac{2}{4} * \log_2\left(\frac{2}{4}\right)\right] - \left[\frac{2}{4} * \log_2\left(\frac{2}{4}\right)\right] = \boxed{1}$

3b) $E_{mild} = -\left[\frac{4}{6} * \log_2\left(\frac{4}{6}\right)\right] - \left[\frac{2}{6} * \log_2\left(\frac{2}{6}\right)\right] = \boxed{0.91829}$

3c) $E_{cool} = -\left[\frac{3}{4} * \log_2\left(\frac{3}{4}\right)\right] - \left[\frac{1}{4} * \log_2\left(\frac{1}{4}\right)\right] = \boxed{0.811278}$

$$Gain(S_{total}, Temp) = 0.94029 - \left[\frac{4}{14} * 1 + \frac{6}{14} * 0.918 + \frac{4}{14} * 0.8112\right]$$

$$= \boxed{0.02922}$$

4) $\boxed{Gain(S_{total}, Humidity) = ?}$

Humidity Value → High, Normal

4a) $E_{high} = -\left[\frac{3}{7} * \log_2\left(\frac{3}{7}\right)\right] - \left[\frac{4}{7} * \log_2\left(\frac{4}{7}\right)\right] = 0.524 + 0.461 = \boxed{0.985}$

4b) $E_{normal} = -\left[\frac{6}{7} * \log_2\left(\frac{6}{7}\right)\right] - \left[\frac{1}{7} * \log_2\left(\frac{1}{7}\right)\right] = 0.191 + 0.401 = \boxed{0.592}$

$$Gain(S_{total}, Humidity) = 0.94029 - \left[\frac{7}{14} * 0.985 + \frac{7}{14} * 0.592\right] = \boxed{0.152}$$

Scanned with CamScanner

5) $\boxed{Gain(S_{total}, Wind) = ?}$ Wind Values → Weak, → Strong

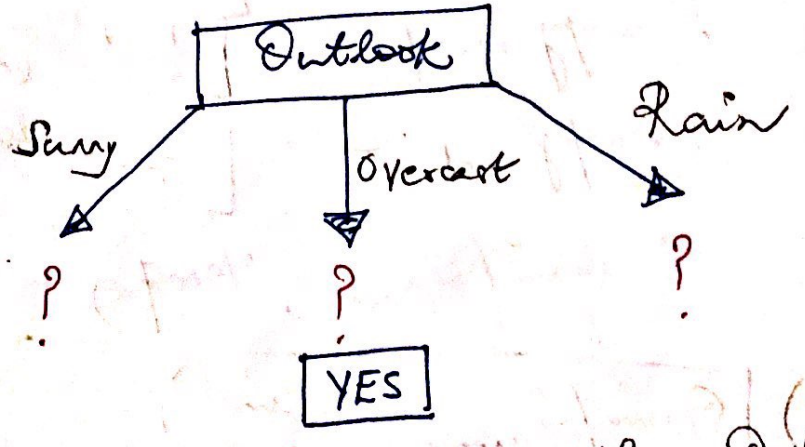
5a) $E_{weak} = -\left[\frac{6}{8} * \log_2\left(\frac{6}{8}\right)\right] - \left[\frac{2}{8} * \log_2\left(\frac{2}{8}\right)\right] = 0.311 + 0.5 = \boxed{0.811}$

5b) $E_{strong} = -\left[\frac{3}{6} * \log_2\left(\frac{3}{6}\right)\right] - \left[\frac{3}{6} * \log_2\left(\frac{3}{6}\right)\right] = 0.5 + 0.5 = \boxed{1}$

$Gain(S_{total}, Wind) = 0.94029 - \left[\left(\frac{8}{14} * 0.811\right) + \left(\frac{6}{14} * 1\right)\right] = 0.94029 - 0.891$
$= \boxed{0.048}$

6) Select the Feature with highest Info Gain as Root Node.

Feature	Info Gain
Outlook	0.24675 ✓
Temp	0.02922
Humidity	0.152
Wind	0.048



7) Look at the Probability Table to see if there are any decisions that can be made, where either samples of 'YES' or 'NO' decisions, don't exist.

Root Node Feature Value	YES	NO
Outlook = Overcast	4	0

Inference: When Outlook is Overcast, Decision is to Play Tennis, since no negative samples exist.

Scanned with CamScanner

8) Find the Entropy & Info gain for all Outlook = Sunny samples.

NOTE: It is no longer $Entropy_{total} = 0.94029$ for 14 samples.

It is $Entropy_{Sunny} = 0.97095$ for 5 samples of 'Sunny'

(Refer Page 1)

8a) $\boxed{Gain(S_{Sunny}, Temp) = ?}$

Samples of 'Sunny'

$$Entropy_{hot} = -\left[\frac{0}{2} * \log_2\left(\frac{0}{2}\right)\right] - \left[\frac{2}{2} * \log_2\left(\frac{2}{2}\right)\right] = |0| = \boxed{0}$$

$$Entropy_{mild} = -\left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] - \left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] = |1| = \boxed{1}$$

$$Entropy_{cool} = -\left[\frac{1}{1} * \log_2\left(\frac{1}{1}\right)\right] = \left[\frac{0}{1} * \log_2\left(\frac{0}{1}\right)\right] = |0| = \boxed{0}$$

$$Gain(S_{Sunny}, Temp) = 0.97095 - \left[\left(\frac{2}{5} * 0\right) + \left(\frac{2}{5} * 1\right) + \left(\frac{1}{5} * 0\right)\right] = \boxed{0.571}$$

8b) $\boxed{Gain(S_{Sunny}, Wind) = ?}$

$$Entropy_{weak} = \left[\frac{1}{3} * \log_2\left(\frac{1}{3}\right)\right] - \left[\frac{2}{3} * \log_2\left(\frac{2}{3}\right)\right] = 0.528 + 0.389 = \boxed{0.917}$$

$$Entropy_{strong} = \left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] - \left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] = \boxed{1}$$

$$Gain(S_{Sunny}, Wind) = 0.9705 - \left[\left(\frac{3}{5} * .917\right) + \left(\frac{2}{5} * 1\right)\right] = \boxed{0.0205}$$

Scanned with CamScanner

8c) $\boxed{Gain(S_{Sunny}, Humidity) = ?}$

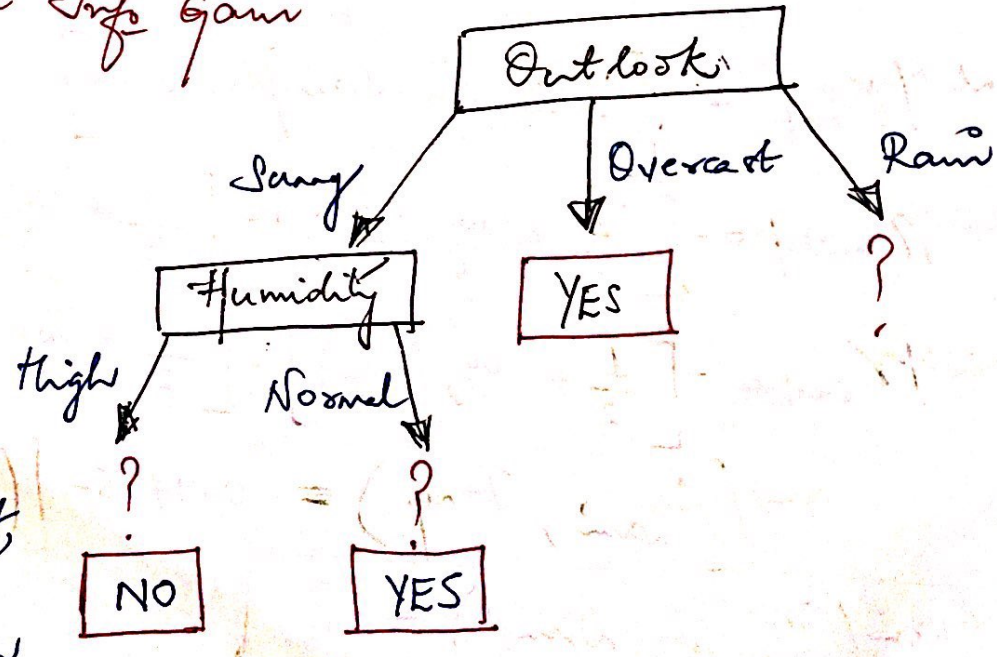
$$Entropy_{high} = \left[\frac{0}{3} * \log_2\left(\frac{0}{3}\right)\right] - \left[\frac{3}{3} * \log_2\left(\frac{3}{3}\right)\right] = \boxed{0}$$

$$Entropy_{normal} = \left[\frac{2}{2} * \log_2\left(\frac{2}{2}\right)\right] - \left[\frac{0}{2} * \log_2\left(\frac{0}{2}\right)\right] = \boxed{0}$$

$$Gain(S_{Sunny}, Humidity) = 0.97095 - \left[\left(\frac{3}{5} * 0\right) + \left(\frac{2}{5} * 0\right)\right] = \boxed{0.97095}$$

9) Select the Feature with highest Info Gain

Feature	Gain
Temp	0.571
Wind	0.0205
Humidity	0.9709 ✓



10) Are there samples of Outlook = Sunny, Humidity = high, Humidity = Normal where 'YES' or 'No' decisions can be readily made as negative or positive decisions samples don't exist

	YES	NO
Outlook = Sunny, Humidity = high	0	3
Outlook = Sunny, Humidity = Normal	2	0

Scanned with CamScanner

11) Find Entropy & Info gain for all Outlook = Rainy samples

NOTE: $Entropy_{Rainy} = 0.9705$ (Refer Page 1) for 5 samples of Rain

11a) $\boxed{Gain(S_{Rain}, Temp) = ?}$

S_{Rain} → Samples of 'Rain'

$Entropy_{hot} = [\longrightarrow \text{No samples} \longleftarrow]$

$Entropy_{mild} = -\left[\frac{2}{3} * \log_2\left(\frac{2}{3}\right)\right] - \left[\frac{1}{3} * \log_2\left(\frac{1}{3}\right)\right] = \left|-(0.528 + 0.389)\right| = \boxed{0.918}$

$Entropy_{cool} = -\left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] - \left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] = \boxed{1}$

$Gain(S_{Rain}, Temp) = 0.9705 - \left[\left(\frac{3}{5} * 0.918\right) + \left(\frac{2}{5} * 1\right)\right] = \boxed{0.019}$

11b) $\boxed{Gain(S_{Rain}, Wind) = ?}$

$Entropy_{weak} = -\left[\frac{3}{3} * \log_2\left(\frac{3}{3}\right)\right] - \left[\frac{0}{3} * \log_2\left(\frac{0}{3}\right)\right] = \boxed{0}$

$Entropy_{strong} = -\left[\frac{0}{2} * \log_2\left(\frac{0}{2}\right)\right] - \left[\frac{2}{2} * \log_2\left(\frac{2}{2}\right)\right] = \boxed{0}$

$Gain(S_{Rain}, Wind) = \boxed{0.9705}$

Scanned with CamScanner

11 c) $\boxed{Gini(S_{Rain}, Humidity) = ?}$

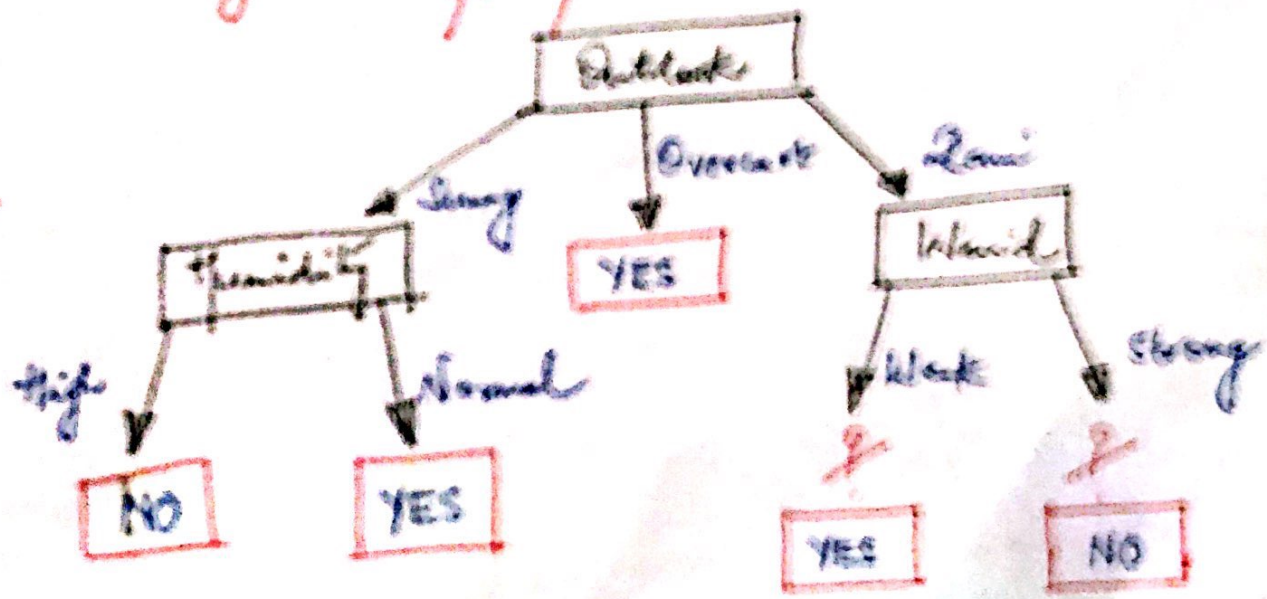
$$Entropy_{high} = -\left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] - \left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] = \boxed{1}$$

$$Entropy_{normal} = -\left[\frac{2}{3} * \log_2\left(\frac{2}{3}\right)\right] - \left[\frac{1}{3} * \log_2\left(\frac{1}{3}\right)\right] = \boxed{0.918}$$

$$Gini(S_{Rain}, Humidity) = 0.9705 - \left[\left(\frac{2}{5} * 1\right) + \left(\frac{3}{5} * 0.918\right)\right] = \boxed{0.019}$$

12) Select the feature with the highest info Gain

Feature	Gain
Temp	0.019
Wind	0.9705 ✓
Humidity	0.019



13)

Outlook = Rain	YES	NO
Wind = Weak	3	0
Wind = Strong	0	2

Find ID3 Decision Tree

Scanned with CamScanner

Classification Tree Algo – Numerical

* Refer Probability Table of ID3 Algo. (Page 1)
* Calculate GINI INDEX of each Feature

$$\text{GINI Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

(Attribute = Value)

$$\text{GINI Index (Attribute)} = \sum_{V = \text{Values}} P_v * GI(v)$$

1) $GINI(\text{Outlook}, \text{'Sunny'}) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right]$

$= 1 - [0.15 + 0.36] = \boxed{0.48}$

$GINI(\text{Outlook}, \text{'Overcast'}) = 1 - \left[\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right] = 1 - 1 = \boxed{0}$

$GINI(\text{Outlook}, \text{'Rain'}) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 1 - [0.36 + 0.15] = \boxed{0.48}$

$GINI(\text{'Outlook'}) = \frac{5}{14} * 0.48 + \frac{4}{14} * 0 + \frac{5}{14} * 0.48 = 0.171 + 0 + 0.171 = \boxed{0.342}$

2) $GINI(\text{Temp}, \text{'Hot'}) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 1 - [0.25 + 0.25] = \boxed{0.5}$

$GINI(\text{Temp}, \text{'Mild'}) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 1 - [0.44 + 0.11] = \boxed{0.45}$

$GINI(\text{Temp}, \text{'Cool'}) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 1 - [0.563 + 0.063] = \boxed{0.374}$

$GINI(\text{'Temp'}) = \frac{4}{14} * 0.5 + \frac{6}{14} * 0.45 + \frac{4}{14} * 0.374 = 0.143 + 0.193 + 0.107$

$= \boxed{0.443}$

Scanned with CamScanner

3) $GINI(\text{Humidity}, \text{'High'}) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 1 - [0.183 + 0.326] = \boxed{0.491}$

$GINI(\text{Humidity}, \text{'Normal'}) = 1 - \left[\left(\frac{6}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right] = 1 - [0.735 + 0.02] = \boxed{0.245}$

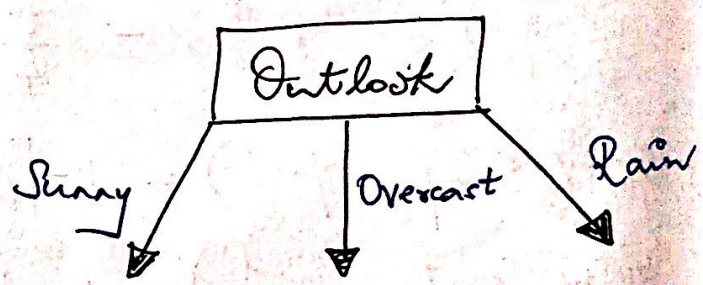
$GINI(\text{Humidity}) = \left(\frac{7}{14} * 0.491\right) + \left(\frac{7}{14} + 0.245\right) = 0.2455 + 0.122 = \boxed{0.368}$

4) $GINI(\text{Wind}, \text{'Weak'}) = 1 - \left[\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right] = 1 - [0.563 + 0.063] = \boxed{0.375}$

$GINI(\text{Wind}, \text{'Strong'}) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 1 - [0.25 + 0.25] = \boxed{0.5}$

$GINI(\text{Wind}) = \left(\frac{8}{14} * 0.375\right) + \left(\frac{6}{14} * 0.5\right) = 0.214 + 0.214 = \boxed{0.428}$

Feature	GINI Index
Outlook	0.342 ✓
Temp	0.443
Humidity	0.368
Wind	0.428



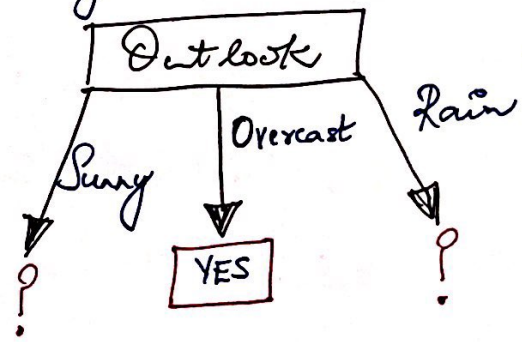
5) Select Root Node of Decision Tree (classification) based on least GINI Index value. Here its the Feature 'Outlook'.

Scanned with CamScanner

6) Analyze the dataset to find out how many 'YES' & 'No' decisions for each feature, Value of Root Node (here 'Outlook')

VALUE OF ROOT NODE	YES	No
SUNNY	2	3
OVERCAST	4	0
RAIN	3	2

Inference: Irrespective of other features ('Wind', 'Temp', 'Humidity') if the Outlook is 'Overcast', Decision 'YES' to Play Tennis.



7) Calculate GINI INDEX for each feature for the Root feature values. Here 'Outlook' values.

→ SUNNY
→ RAIN

a) GINI(Outlook = Sunny, 'Temp') = ?

$$GINI(Outlook = 'Sunny', Temp = 'Hot') = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right]$$
$$= 1 - 1 = \boxed{0}$$

$$GINI(Outlook = 'Sunny', Temp = 'Mild') = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$
$$= 1 - 0.25 - 0.25 = \boxed{0.5}$$

$$GINI(Outlook = Sunny, Temp = Cool) = 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right]$$
$$= 1 - 1 = \boxed{0}$$

$$GINI(Outlook = Sunny, Temp) = \left(\frac{2}{5} * 0\right) + \left(\frac{2}{5} * 0.5\right) + \left(\frac{1}{5} * 0\right) = \boxed{0.2}$$

Scanned with CamScanner

b) Gini(Outlook = 'Sunny', Humidity) = ?

Gini(Outlook = 'Sunny', Humidity = 'High') = $1-\left[\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right] = 1-1 = \boxed{0}$

Gini(Outlook = 'Sunny', Humidity = 'Normal') = $1-\left[\left(\frac{2}{2}\right)^2+\left(\frac{0}{2}\right)^2\right] = 1-1 = \boxed{0}$

Gini(Outlook = 'Sunny', Humidity) = $\left(\frac{3}{5}\times 0\right)+\left(\frac{2}{5}\times 0\right) = \boxed{0}$ ← STOP CONDITION

c) Gini(Outlook = 'Sunny', Wind) = ?

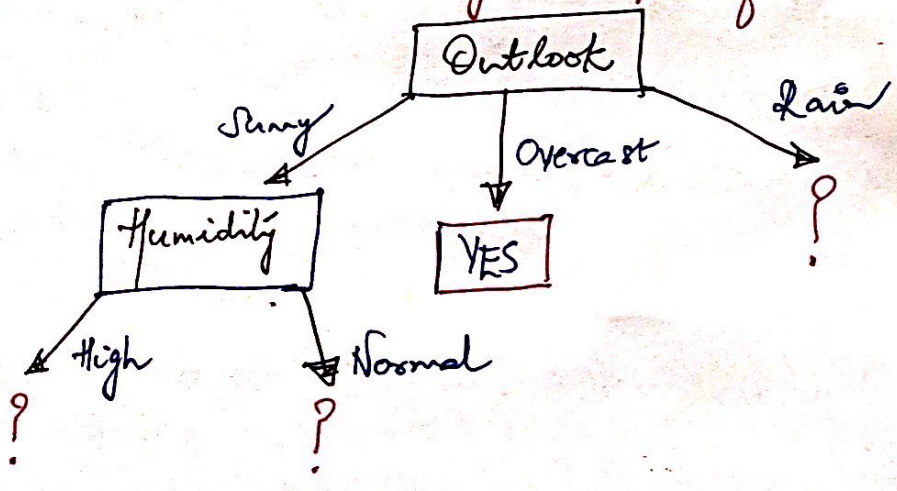
Gini(Outlook = 'Sunny', Wind = Weak) = $1-\left[\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right] = 1-0.734 = \boxed{0.266}$

Gini(Outlook = 'Sunny', Wind = strong) = $1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right] = 1-0.8 = \boxed{0.2}$

Gini(Outlook = Sunny, Wind) = $\left(\frac{3}{5}\times 0.266\right)+\left(\frac{2}{5}\times 0.2\right) = \boxed{0.466}$

8) Find the feature with lowest Gini index in 'Sunny' sample of 'Outlook'

Feature	Gini index
Temp	0.2
Humidity	0 ✓
Wind	0.466



Scanned with CamScanner

9) Calculate GINI INDEX for the left out feature value of Outlook, here 'Rain'

a) GINI(Outlook = 'Rainy', Temp) = ?

GINI(Outlook = Rainy, Temp = 'Hot') = $\boxed{0}$ No samples

GINI(Outlook = Rainy, Temp = 'Cool') = $1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right] = 1-0.5 = \boxed{0.5}$

GINI(Outlook = Rain, Temp = 'Mild') = $1-\left[\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right] = 1-0.556 = \boxed{0.44}$

GINI(Outlook = Rain, Temp) = $\left(\frac{2}{5} * 0.5\right)+\left(\frac{3}{5} * 0.44\right) = \boxed{0.46}$

b) GINI(Outlook = 'Rainy', Wind) = ?

GINI(Outlook = Rain, Wind = 'weak') = $1-\left[\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2\right] = 1-1 = \boxed{0}$

GINI(Outlook = Rain, Wind = 'strong') = $1-\left[\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2\right] = 1-1 = \boxed{0}$

GINI(Outlook = Rain, Wind) = $\left(\frac{3}{5} * 0\right)+\left(\frac{2}{5} * 0\right) = \boxed{0}$ ← STOP CONDITION

c) GINI(Outlook = 'Rain', Humidity) = ?

GINI(Outlook = Rainy, Humidity = 'High') = $1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right] = 1-0.5 = \boxed{0.5}$

GINI(Outlook = Rain, Humidity = Normal) = $1-\left[\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right] = 1-0.556 = \boxed{0.44}$

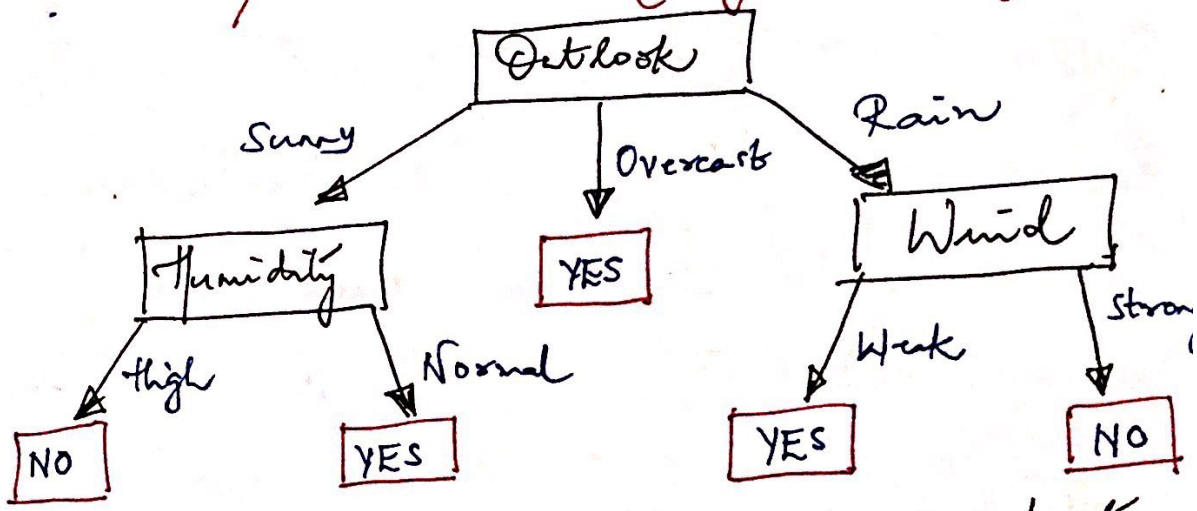
GINI(Outlook = Rain, Humidity) = $\left(\frac{2}{5} * 0.5\right)+\left(\frac{3}{5} * 0.44\right) = \boxed{0.46}$

Scanned with CamScanner

10) Find the feature with the lowest Gini Index in 'Rainy' value of 'Outl

Feature	Gini Index
Temp	0.46
Humidity	0.46
Wind	0 ✓

Final classification Tree →



11) Draw the Decision Tree to complete the path 'Outlook = Rain' to

* depict the Feature 'Wind' as Parent Node. ('Strong')
* depict the path for (i) {Outlook = 'Rainy', Wind = 'Strong', Outlook = 'Rainy', Humidity = High} (ii) {Outlook = Sunny, Wind = 'Weak', Outlook = Sunny, Humidity = Normal} by

– Counting the samples for (i) & (ii).

VALUE OF PARENT NODE	YES	NO
HUMIDITY		
High	0	3
Normal	2	0

VALUE OF PARENT NODE	YES	NO
WIND		
Weak	3	0
Strong	0	2

12) Inference: Since the Gini Index values of (Outlook = 'Sunny', Humidity) (Outlook = 'Rainy', Wind) are zero the STOP CONDITION of decision tree is reached.

⇒ Leaf Node Decisions are reached.

Scanned with CamScanner