Data Science and Artificial Intelligence

Machine Learning

Decision Tree

Lecture No. 1

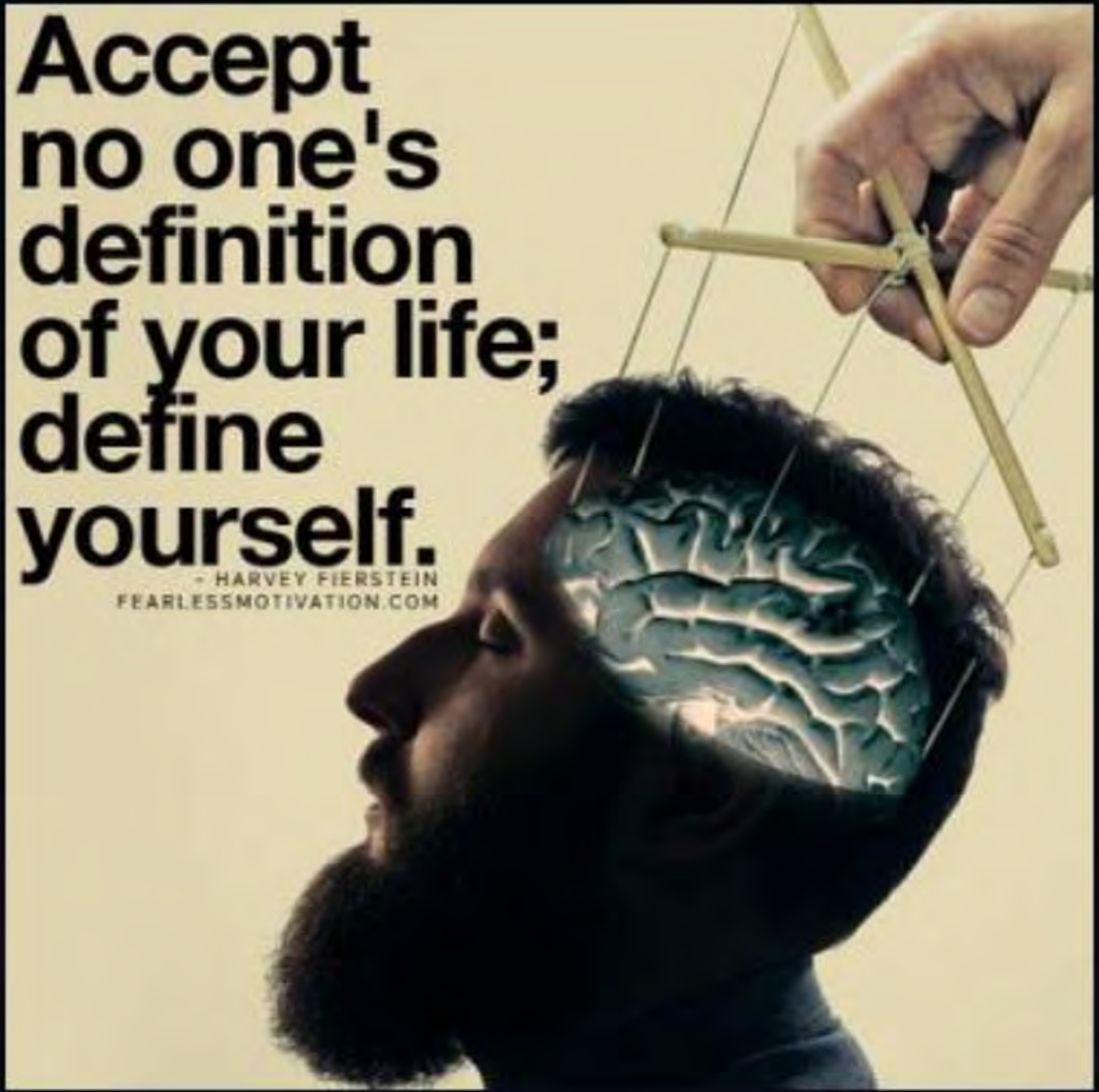
By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — decisiontree
- Topic — Basic
- Topic
- Topic
- Topic

Topics to be Covered

- **Topic** Gini Impurity
- **Topic** Entropy
- **Topic** Variance
- **Topic** Info. Gain
- **Topic**


Accept no one's definition of your life; define yourself.
- HARVEY FIERSTEIN
FEARLESSMOTIVATION.COM

Decision Trees → non parametric

• non linear

→ prone to overfit

→ Start with Root node, keep on grouping/splitting ⇒ to get homogeneous node

→ To reduce confusion

→ How to decide on new test point

REVISION

Why do we split/group ?

done

REVISION

How final decision is made.

done

@leaf node

y_p = majority
or
avg.

data ⇒ class 1
class 2

80% class1 ← oversampling

20% class2 ← undersampling

assymmetric data

Decision Tree

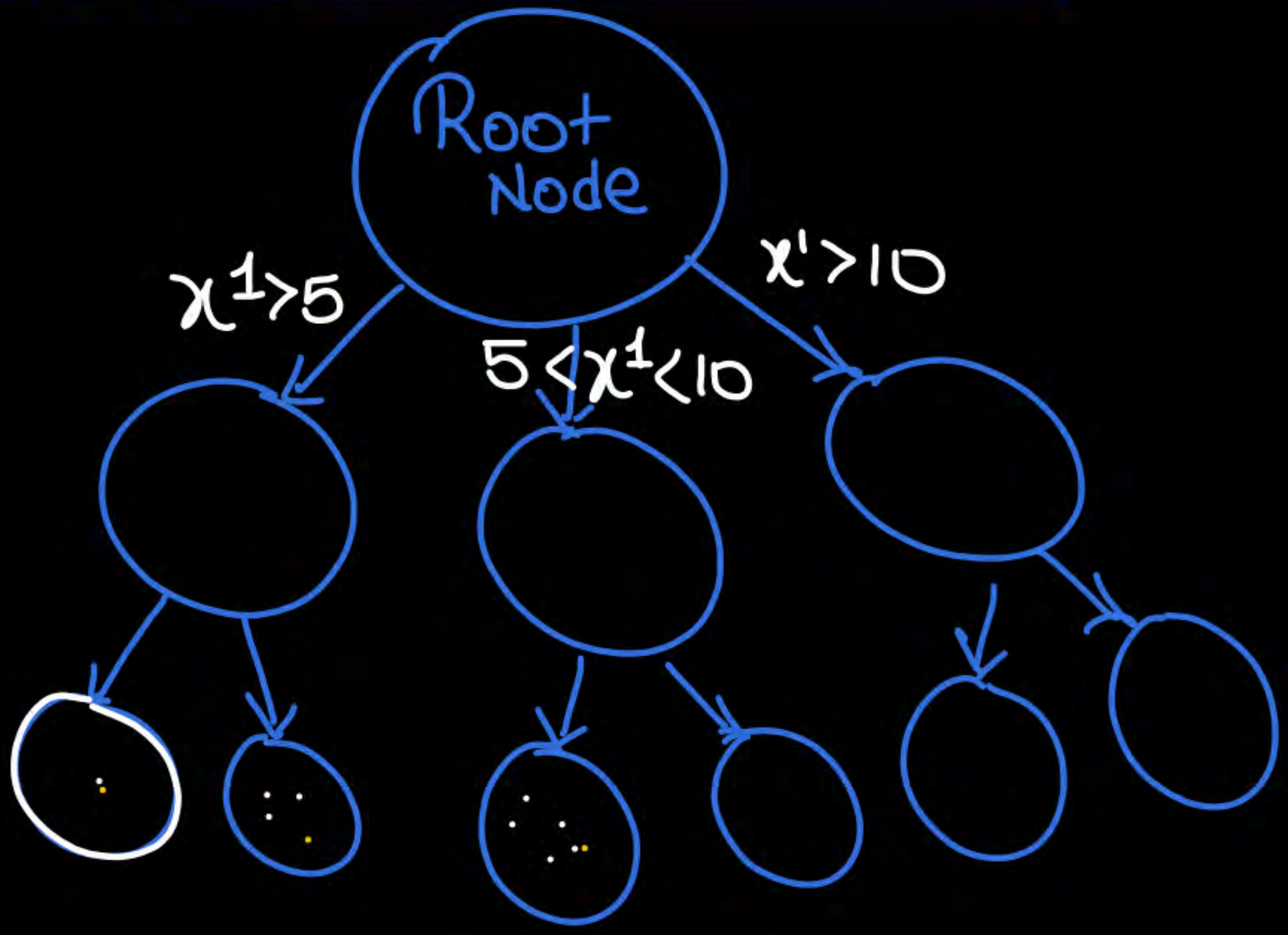
What is a Decision Tree

We start with the full training data at the root node

Now Based on some variable the whole input space divided

Keep dividing the input space till you reach the final [stopping criteria]

Or you reach the stopping condition



What is a Decision Tree

So each leaf node will show
Some region (2D Case)

$$\begin{cases} \text{if } (x^1 < 5, x^2 < 0) \rightarrow y_p = \text{avg} \\ \qquad\qquad\qquad\qquad\quad = \text{max class} \\ \text{if } (5 < x^1 < 10, x^2 > 0) \rightarrow y_p = \text{avg} \\ \qquad\qquad\qquad\qquad\quad = \text{max clas} \\ \vdots \end{cases}$$

How this decision tree is stored in memory...

⇒ Space Complexity is very less becozon if/else prog. is to be stored

⇒ Testing time Complexity is also very less.

- But generating decisiontree is very Complex.

- KNN is effected by outlier → Y/N.

  No **

Decision Tree

What is a Decision Tree

Some Important Terms
Decision Tree is Non Linear

Decision tree is Non parametric, and non linear

done

Because here we make no assumption on the pattern of the data, we simply take and data and work on it.

Decision Tree

Lets see an example

Day	Outlook	Temperature	Humidity	Wind	Play Golf
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

We have to predict whether we have to play or not... Classification Problem

At root node we can choose any attribute for decision

- Classification Problem

- So to create the decision tree we have to know that which dimension has to be used for splitting and how to split.

- In decision tree the algorithm check all the dimension and then find the best how

Decision Tree

How to select the attribute for splitting ?

We do splitting to reduce the confusion, after splitting we need most homogeneous nodes. Where the concentration of a particular label is very high

We can see that in the staring at the root node we have points with Y/N both... hence there was lot of confusion.
i.e. in whole input space we have all the points we need to divide the input space to get regions of similar points

Decision Tree

Which attribute to choose for decision ?

Decision tree is a greedy approach where we check all the possibilities of split and find the best for us

Attribute Selection Measures ...

How to select the attribute for splitting ?

How we select the attribute for splitting ??

Attribute Selection measure ...

How to select the attribute for splitting ?

How to measure node impurity/ node purity/ node homogeneity/ degree of randomness...

Attribute Selection measure ...

How to select the attribute for splitting ?

For classification case : Gini Index and Entropy

For Regression : Variance.

Information Gain : After splitting we measure the reduction in the impurity...

How to select the attribute for splitting ?

Gini Index (for classification)

Concept is if Probability for misclassify is high then Gini Index is high else it is low...

Target in Decision tree is to get $\underbrace{\text{homogeneous node}}_{\text{Pure node}}$.

So to check the impurity of a node we use

$$GI \Rightarrow 1 - \sum_{i=1}^{c} (P_i)^2$$

5 class data

node

20 points

10 → class1 ✓

8 → class2

1 → class3

1 → class4

0 → class5

impure

non homogeneous

$$GI = 1 - \sum_{i=1}^{5} P_i^2$$

$$= 1 - \left(P_1^2 + P_2^2 + P_3^2 + P_4^2 + P_5^2\right)$$

$$= 1 - \left(\left(\frac{10}{20}\right)^2 + \left(\frac{8}{20}\right)^2 + \left(\frac{1}{20}\right)^2 + \left(\frac{1}{20}\right)^2 + 0\right)$$

$$GI = .585$$

5 class

node

20 point
0 → class 1
0 → class 2
20 → class 3
0 → class 4
0 → class 5

Perfect homo →

$$GI = 1 - \sum_{i=1}^{5} P_i^2$$

$$= 1 - \left(P_1^2 + P_2^2 + P_3^2 + P_4^2 + P_5^2\right)$$

$$= 1 - \left(0 + 0 + 1 + 0 + 0\right)$$

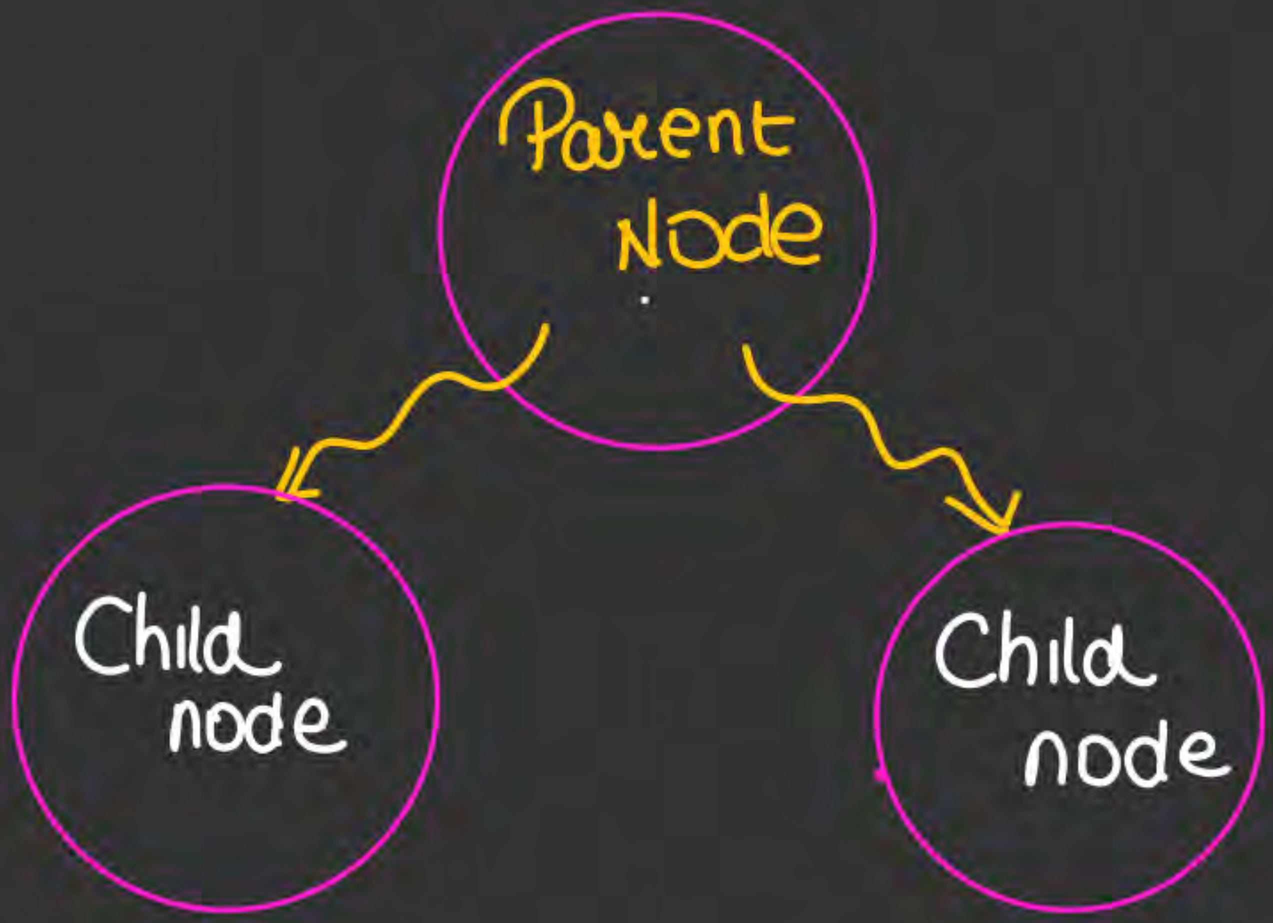
$$= 0$$

GI index $\Rightarrow 1 - \sum_{i=1}^{c} P_i^2$

→ always have value 0 to 1

→ 0 ⇒ Perfectly homogeneous node

→ 1 ⇒ Perfectly non homogeneous node



- Splitting is done to inc homogenity
- So when homogenity inc then GI reduces
- GI Parent > GI children
- Information Gain ⇒ $GI^{Parent} - GI^{Children}$

How to select the attribute for splitting ?

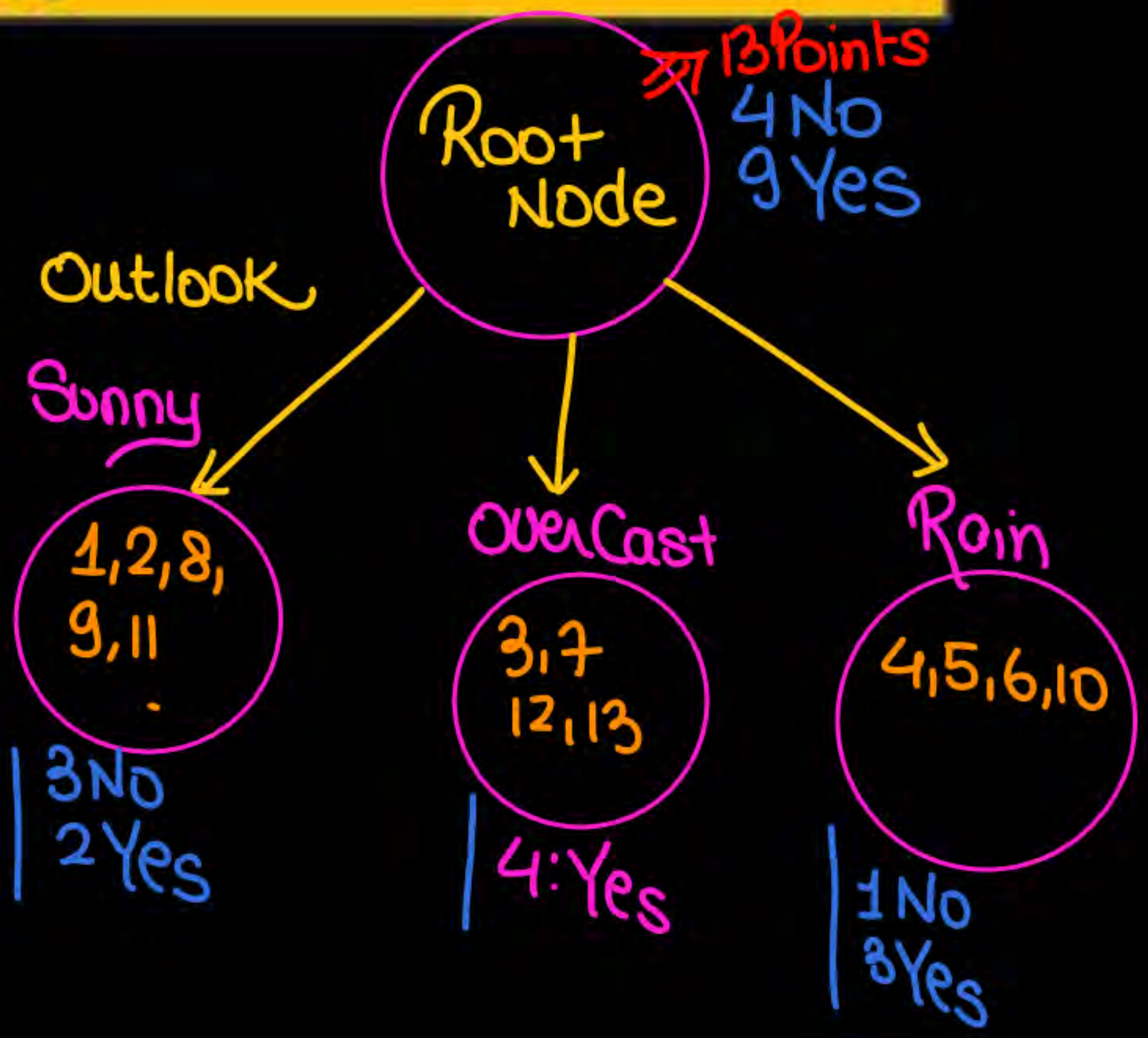
Gini Index (for classification)

We want gini index as low as possible.

Decision Tree

Lets see an example

Day	Outlook	Temperature	Humidity	Wind	Play Golf	
D1	Sunny	Hot	High	Weak	No	1
D2	Sunny	Hot	High	Strong	No	2
D3	Overcast	Hot	High	Weak	Yes	3
D4	Rain	Mild	High	Weak	Yes	4
D5	Rain	Cool	Normal	Weak	Yes	5
D6	Rain	Cool	Normal	Strong	No	6
D7	Overcast	Cool	Normal	Strong	Yes	7
D8	Sunny	Mild	High	Weak	No	8
D9	Sunny	Cool	Normal	Weak	Yes	9
D10	Rain	Mild	Normal	Weak	Yes	10
D11	Sunny	Mild	Normal	Strong	Yes	11
D12	Overcast	Mild	High	Strong	Yes	12
D13	Overcast	Hot	Normal	Weak	Yes	13
D14	~~Rain~~	~~Mild~~	~~High~~	~~Strong~~	~~No~~	



2 class

$$GI_{Root} \Rightarrow 1 - \left[P_Y^2 + P_N^2\right]$$

$$\Rightarrow 1 - \left[\left(\frac{9}{13}\right)^2 + \left(\frac{4}{13}\right)^2\right]$$

$$\Rightarrow .426$$

$$GI_{Sunny} = 1 - \left[P_Y^2 + P_N^2\right]$$

$$= 1 - \left[(2/5)^2 + (3/5)^2\right] =$$

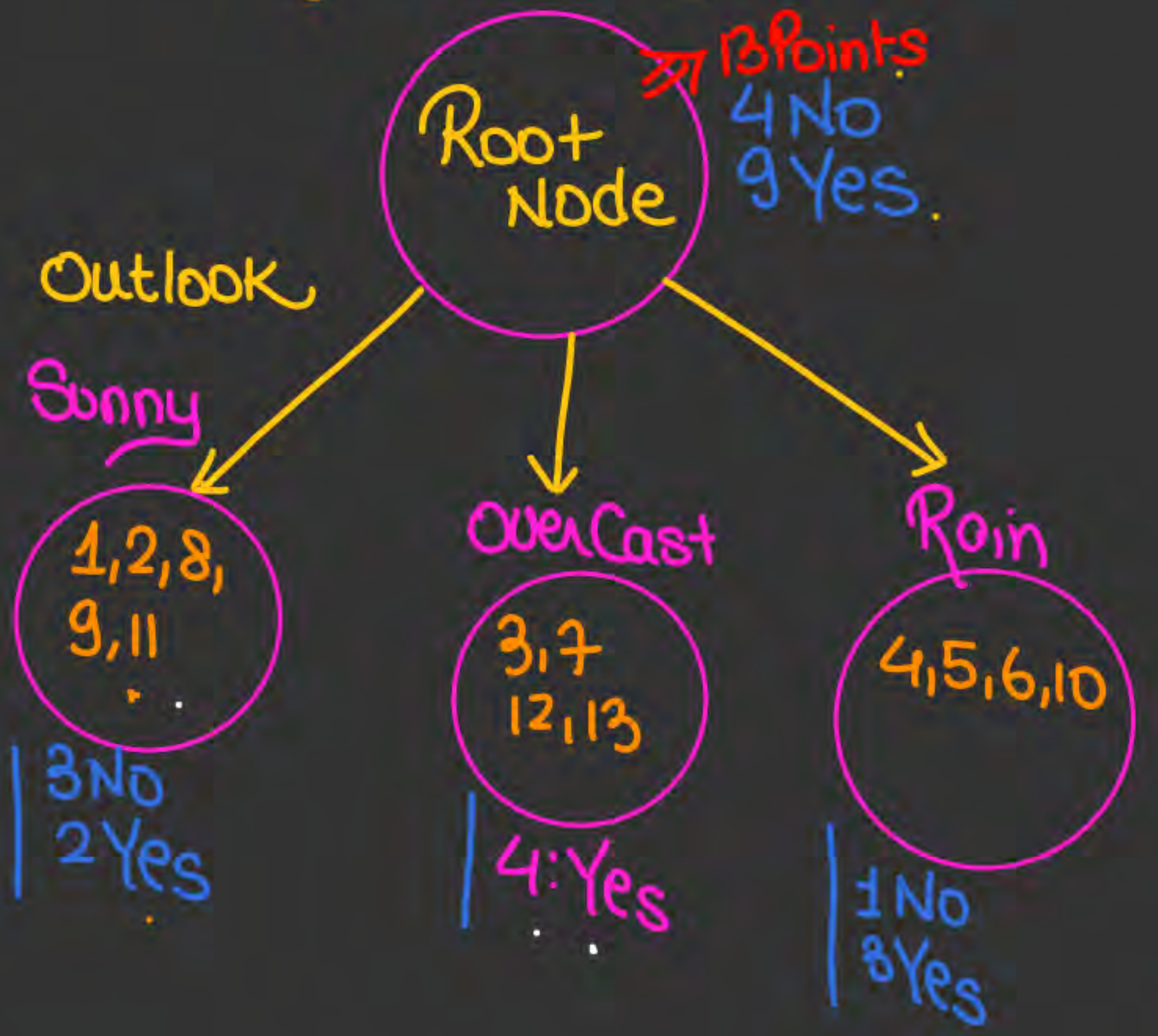
$$= .48$$

$$GI_{overcast} = 1 - \left[P_Y^2 + P_N^2\right]$$

$$= 1 - \left[1 + 0\right] = 0$$

$$GI_{Rain} = 1 - \left[P_Y^2 + P_N^2\right]$$

$$= 1 - \left[(3/4)^2 + (1/4)^2\right] = .375$$



$$IG = (GI^{P} - GI^{C})$$

$$IG = .426 - .3$$
$$= .126$$

GI^{children} = weighted avg of GI of all children

$$\Rightarrow \frac{\sum GI_i \times \text{No of Point in } i^{th} \text{ chid}}{\text{Total No of Points in all children}}$$

$$\Rightarrow \frac{.48 \times 5 + 0 \times 4 + .375 \times 4}{13}$$

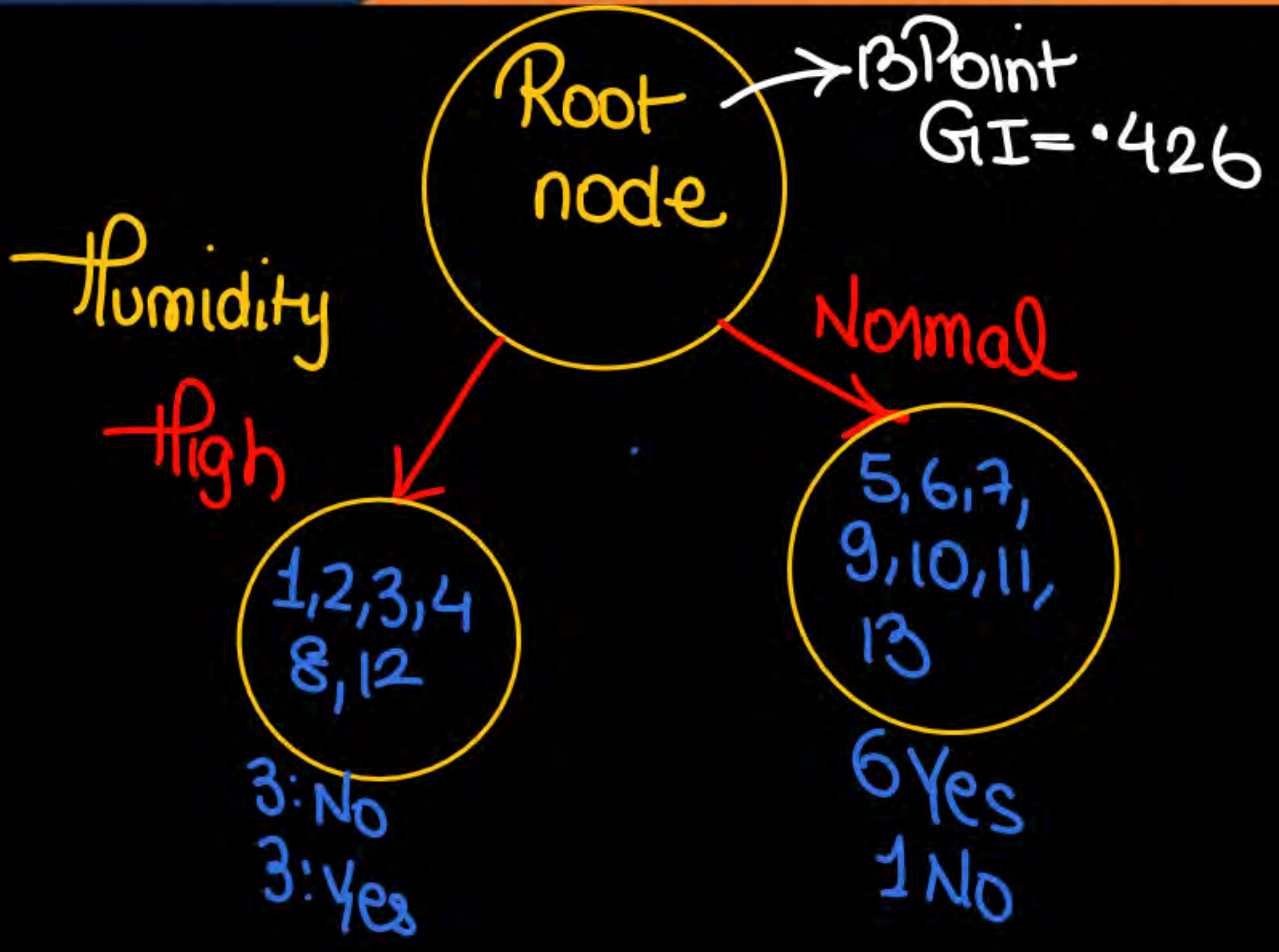
$$\Rightarrow .3$$

Decision Tree

How to select the attribute for splitting ?

Gini Index (for classification)

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No 1
D2	Sunny	Hot	High	Strong	No 2
D3	Overcast	Hot	High	Weak	Yes 3
D4	Rain	Mild	High	Weak	Yes 4
D5	Rain	Cool	Normal	Weak	Yes 5
D6	Rain	Cool	Normal	Strong	No 6
D7	Overcast	Cool	Normal	Strong	Yes 7
D8	Sunny	Mild	High	Weak	No 8
D9	Sunny	Cool	Normal	Weak	Yes 9
D10	Rain	Mild	Normal	Weak	Yes 10
D11	Sunny	Mild	Normal	Strong	Yes 11
D12	Overcast	Mild	High	Strong	Yes 12
D13	Overcast	Hot	Normal	Weak	Yes 13
D14	~~Rain~~	~~Mild~~	~~High~~	~~Strong~~	~~No~~



$$GI^{child} = \frac{1/2 \times 6 + .244 \times 7}{13}$$

$$= .3626$$

$$GI_{high} = 1 - (P_Y^2 + P_N^2)$$

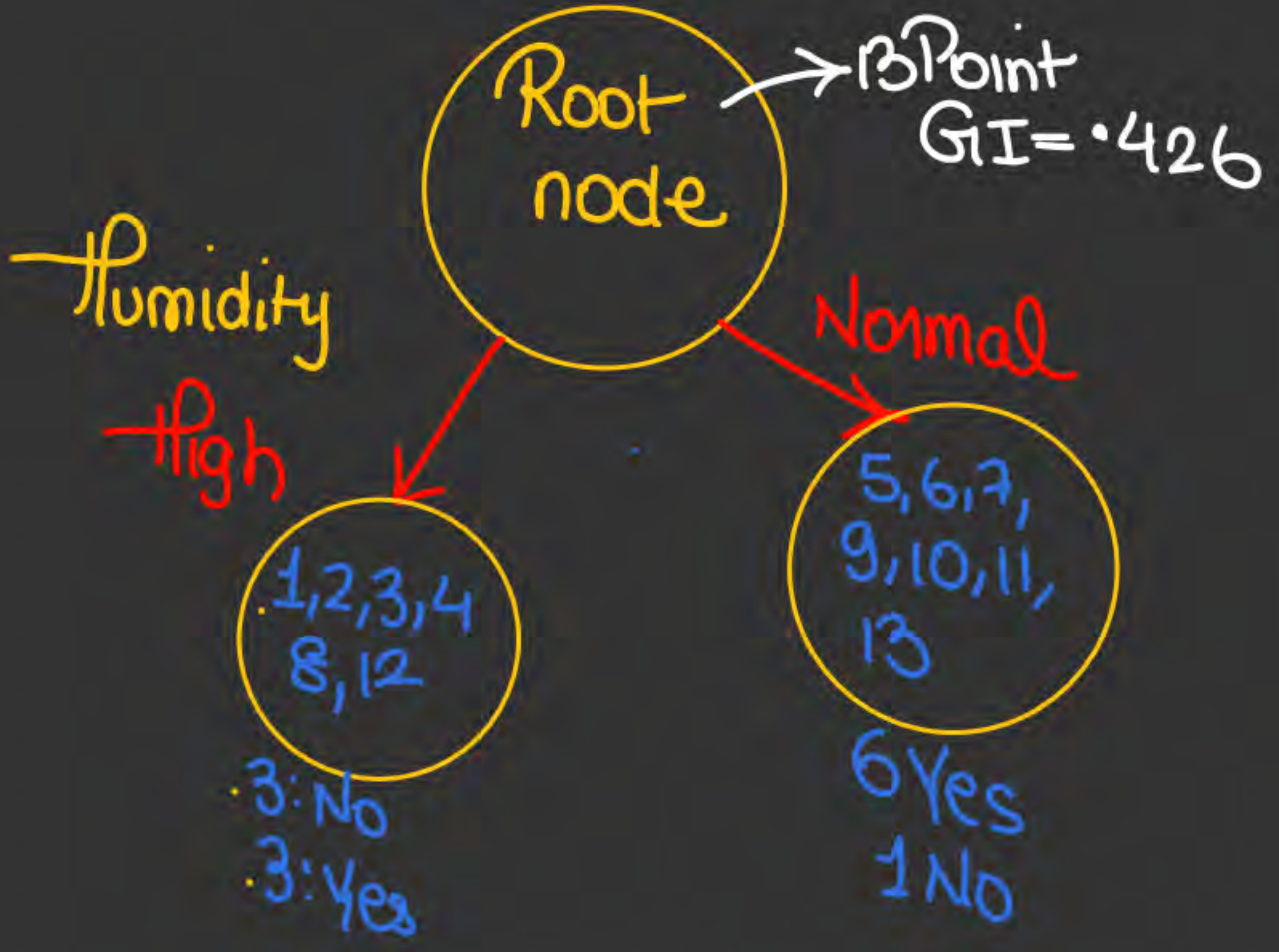
$$= 1 - \left((1/2)^2 + (1/2)^2\right)$$

$$\Rightarrow \frac{1}{2}$$

$$GI_{normal} = 1 - (P_Y^2 + P_N^2)$$

$$= 1 - \left((6/7)^2 + (1/7)^2\right)$$

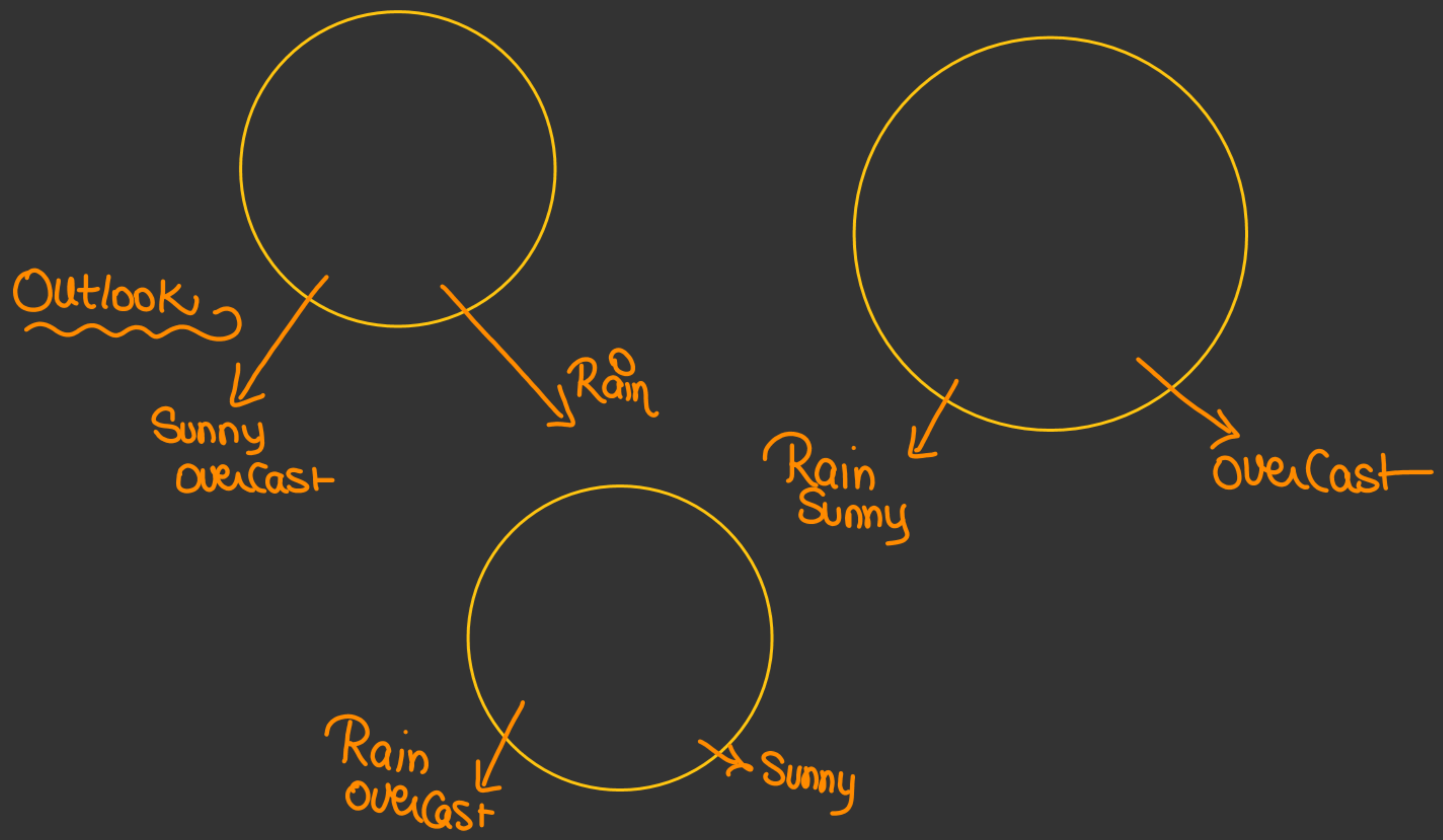
$$= .244$$



$IG = .426 - .362$

$= .06336$

So outlook is better than humid



Outlook
Sunny
overcast
Rain
Rain
Sunny
overcast
Rain
overcast
Sunny

So creating a decision tree is very Complicated

Decision Tree

How to select the attribute for splitting ?

Gini Index (for classification)

It is probability of misclassifying any point in data...

- Probability to misclassify any point ⇒ (If Point is of class1 but predicted that It is of class0 + If Point " " 0 " " "
" " " " 1)

So Probability of misclassification

$P(0/1)$

misclassification
if point is of class 1 but classified as 0
+
if point is of class 0 classified as 1

$P(1/0)$

$$P_{miss} = P(0/1)P(1) + P(1/0)P(0)$$

if 1, 0 are Independent

$$P_{miss} = P_0 P_1 + P_1 P_0$$

$P(E) \Rightarrow P(E/A)P(A) +$
$P(E/B) \cdot P(B) +$
$P(E/C) \cdot P(C) + \ldots$

$P(A/B) \Rightarrow P(A)$
A, B are Independent

Since 2 class ⇒

- $P_0 + P_1 = 1$

$$P_{mis} = P_0 \underbrace{P_1}_{(1-P_0)} + P_1 \underbrace{P_0}_{(1-P_1)}$$

$$= P_0(1-P_0) + P_1(1-P_1)$$

$$= P_0 - P_0^2 + P_1 - P_1^2$$

$$P_{miss} = 1 - P_0^2 - P_1^2$$

$$\Rightarrow GI = 1 - \left(\sum P_i^2\right)$$

- So GI = Probability of misclassification

How to select the attribute for splitting ?

Entropy (for classification)

→ Similar to GI

$$\text{Entropy} \Rightarrow \sum_{i=1}^{c} P_i \log_2 \frac{1}{P_i}$$

$$\Rightarrow - \sum_{i=1}^{c} P_i \log_2 P_i$$

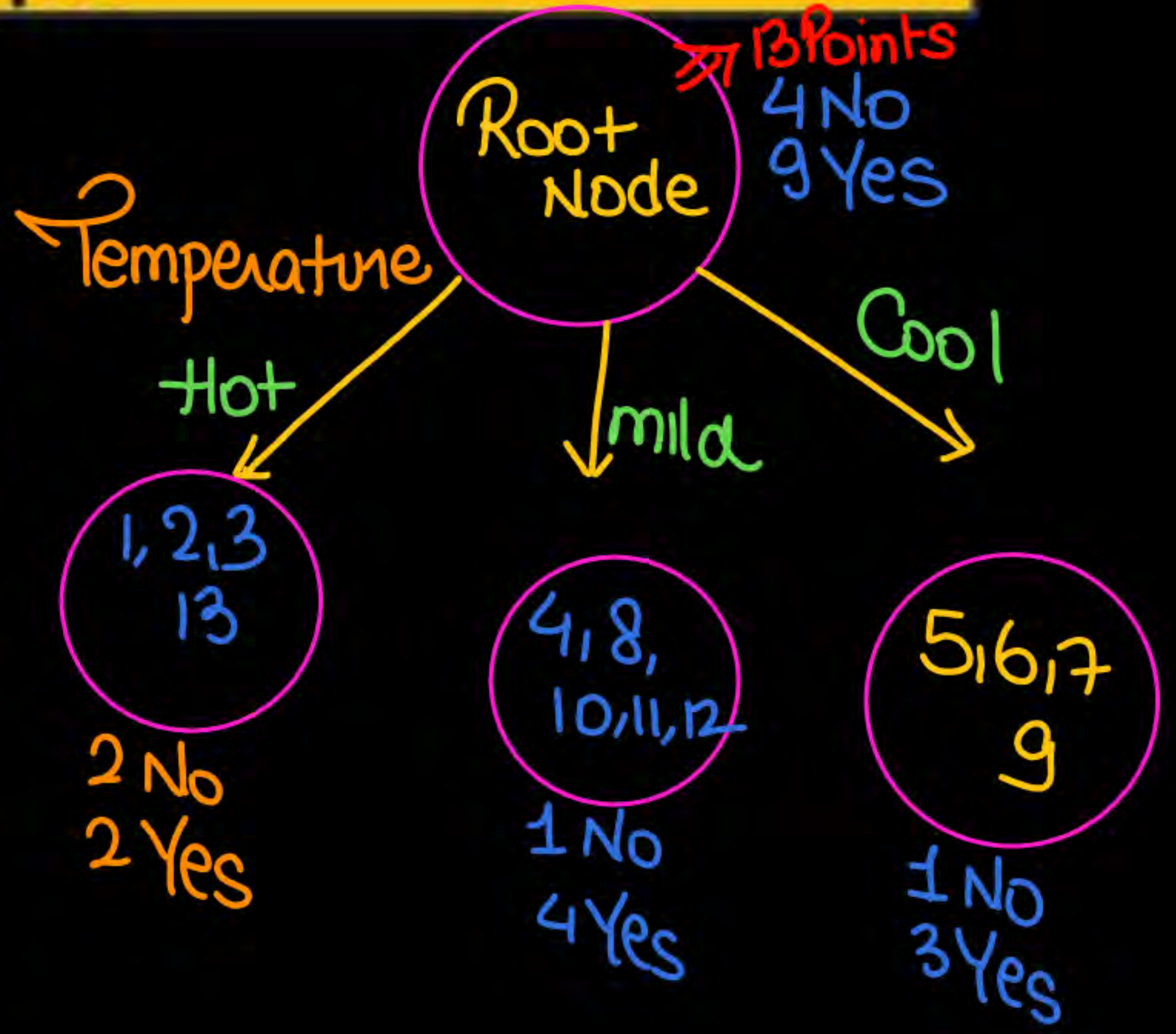
Entropy measure the amount of uncertainty or degree of randomness

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

Decision Tree

Lets see an example

Day	Outlook	Temperature	Humidity	Wind	Play Golf	
D1	Sunny	Hot	High	Weak	No	1
D2	Sunny	Hot	High	Strong	No	2
D3	Overcast	Hot	High	Weak	Yes	3
D4	Rain	Mild	High	Weak	Yes	4
D5	Rain	Cool	Normal	Weak	Yes	5
D6	Rain	Cool	Normal	Strong	No	6
D7	Overcast	Cool	Normal	Strong	Yes	7
D8	Sunny	Mild	High	Weak	No	8
D9	Sunny	Cool	Normal	Weak	Yes	9
D10	Rain	Mild	Normal	Weak	Yes	10
D11	Sunny	Mild	Normal	Strong	Yes	11
D12	Overcast	Mild	High	Strong	Yes	12
D13	Overcast	Hot	Normal	Weak	Yes	13
D14	~~Rain~~	~~Mild~~	~~High~~	~~Strong~~	No	



$$\mathcal{E}_{Root} \Rightarrow P_1 \log_2 \frac{1}{P_1} + P_0 \log_2 \frac{1}{P_0}$$

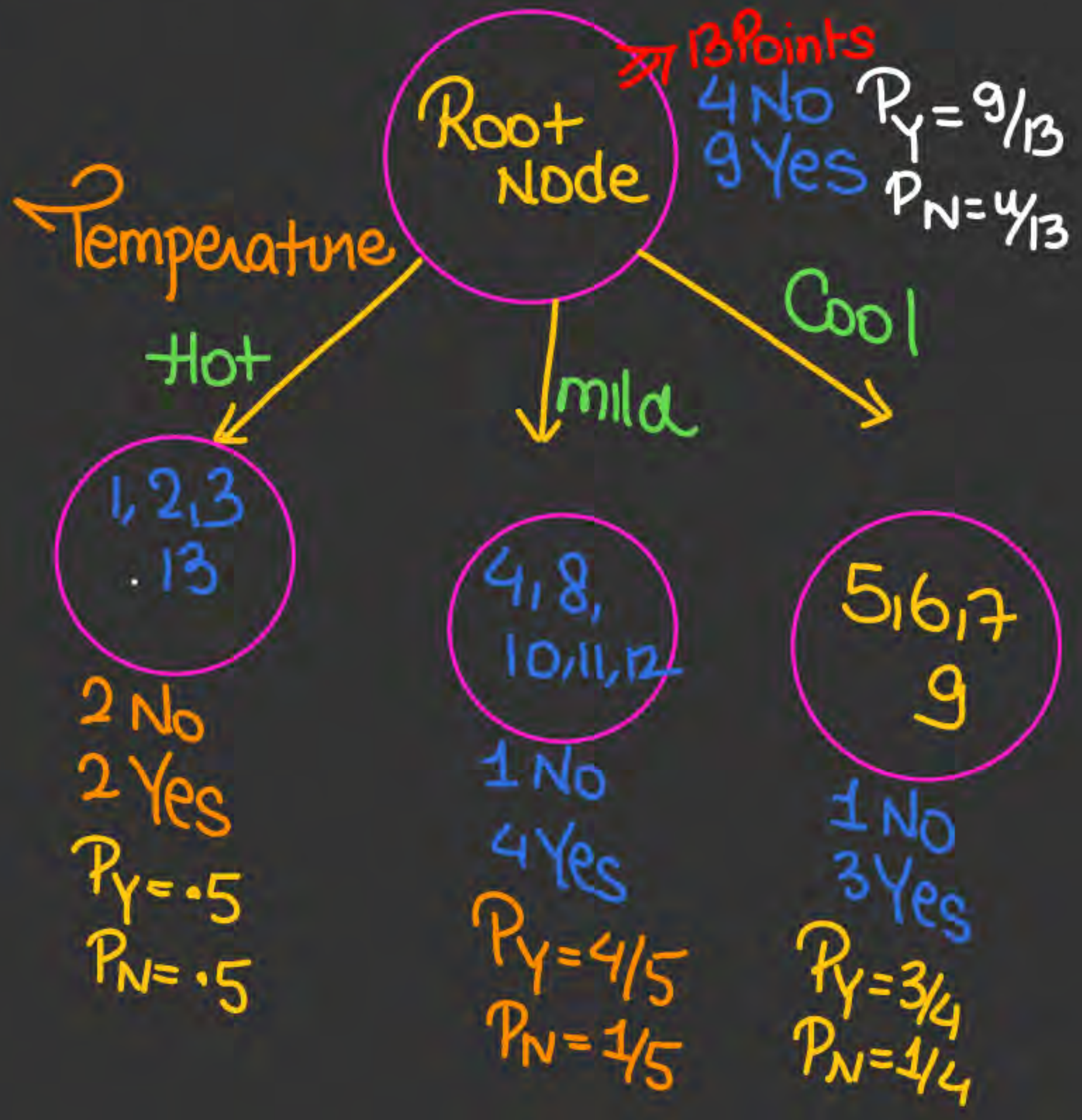
$$\Rightarrow \frac{P_1 \log_{10} \frac{1}{P_1} + P_0 \log_{10} \frac{1}{P_0}}{\log_{10} 2}$$

$$\Rightarrow \frac{\frac{9}{13} \log_{10} 13/9 + \frac{4}{13} \log 13/4}{\log_{10} 2}$$

$$\Rightarrow .8904$$

$$\mathcal{E}_{HOT} \Rightarrow \frac{.5 \log_{10} 2 + .5 \log_{10} 2}{\log_{10} 2} = 1$$

$$\left\{ \log_2 x = \frac{\log_{10} x}{\log_{10} 2} \right\}$$



$$\mathcal{E}_{mild} = \frac{\frac{4}{5}\log_{10} 5/4 + \frac{1}{5}\log_{10} 5}{\log_{10} 2}$$

$$= .721$$

$$\mathcal{E}_{Cool} \Rightarrow \frac{\frac{3}{4}\log 4/3 + \frac{1}{4}\log 4}{\log 2}$$

$$\Rightarrow \underline{.811}.$$

$$\mathcal{E}^{child} \Rightarrow \frac{4\mathcal{E}_h + 5\mathcal{E}_m + 4\mathcal{E}_c}{13}$$

$$\Rightarrow .834$$

$$IG = \mathcal{E}^{Root} - \mathcal{E}^{child}$$

$$= .8904 - .834$$

$$= .055$$

How to select the attribute for splitting ?

Entropy (for classification)

→ GI and entropy measure the non homogenity in node

GI = 0 for Homogeneous node

$0 < GI < 1$

$P_1 = 1$

$P_0 = 0$

5 points
5 → class1
0 → class2

• $\text{Entropy} = 1\log_2 1 + 0\log_2 0$

$= 0$

Decision Tree

How to select the attribute for splitting ?

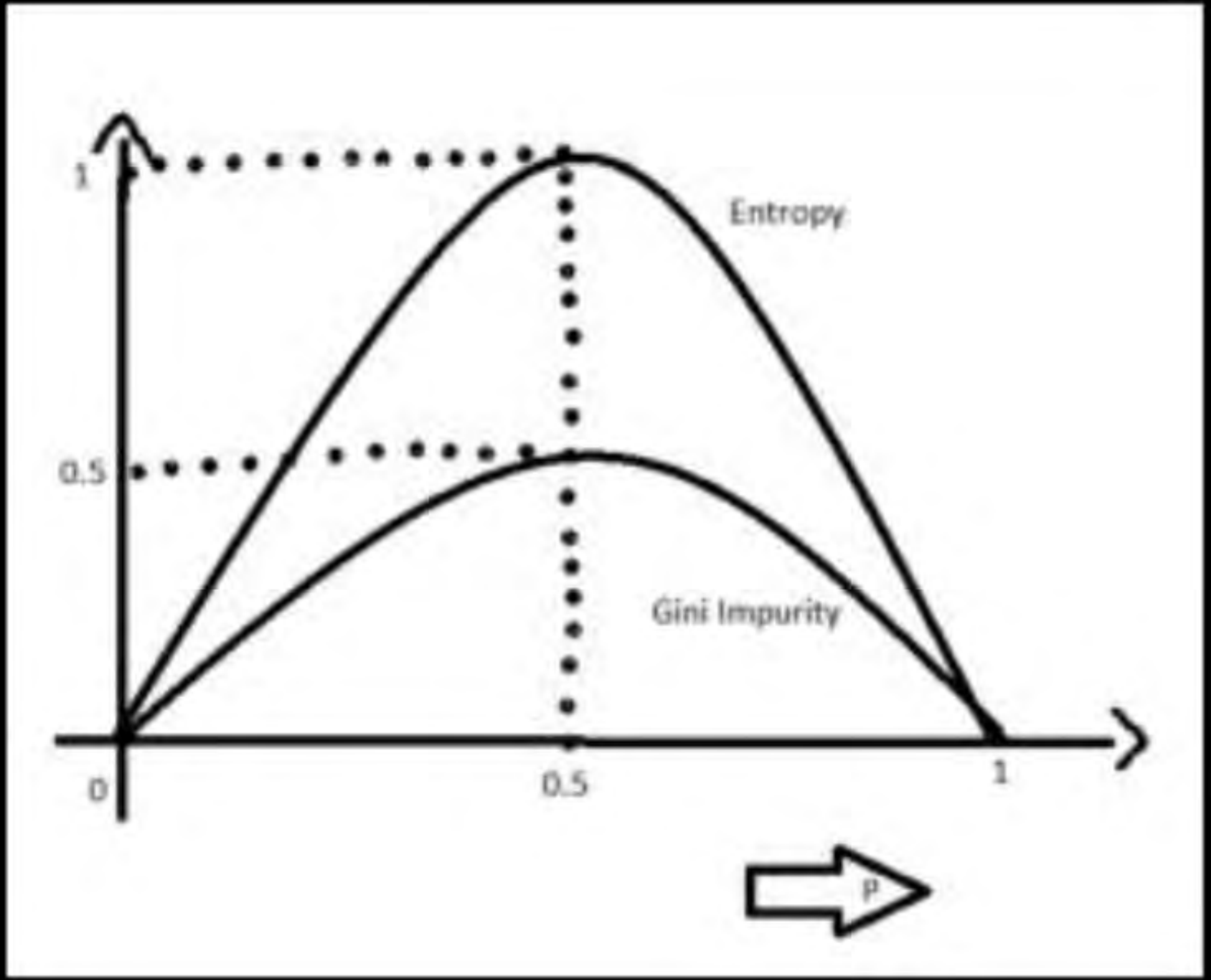
Entropy (for classification)

Entropy=0 in homogeneous Case

more the value of entropy more is Impurity.

max → $\log_2$(No of classes).

Entropy Vs Gini Index



Entropy Vs Gini Index

It is the probability of misclassifying a randomly chosen element in a set.	While entropy measures the amount of uncertainty or randomness in a set.
The range of the Gini index is [0, 1], where 0 indicates perfect purity and 1 indicates maximum impurity.	The range of entropy is [0, log(c)], where c is the number of classes.
Gini index is a linear measure.	Entropy is a logarithmic measure.
It can be interpreted as the expected error rate in a classifier.	It can be interpreted as the average amount of information needed to specify the class of an instance.
It is sensitive to the distribution of classes in a set.	It is sensitive to the number of classes.

How to select the attribute for splitting ?

Entropy Vs Gini Index

It is less robust than entropy.	It is more robust than Gini index.
It is sensitive.	It is comparatively less sensitive.
Formula for the Gini index is Gini(P) = 1 – ∑(Px)^2 , where Pi is the proportion of the instances of class x in a set.	Formula for entropy is Entropy(P) = -∑(Px)log(Px), where pi is the proportion of the instances of class x in a set.

THANK - YOU