DATA SCIENCE

DATE : 30.07.2024

DT/NT :

LESSON : MACHINE LEARNING

SUBJECT : DECISION TREE (CART)

BATCH : 247

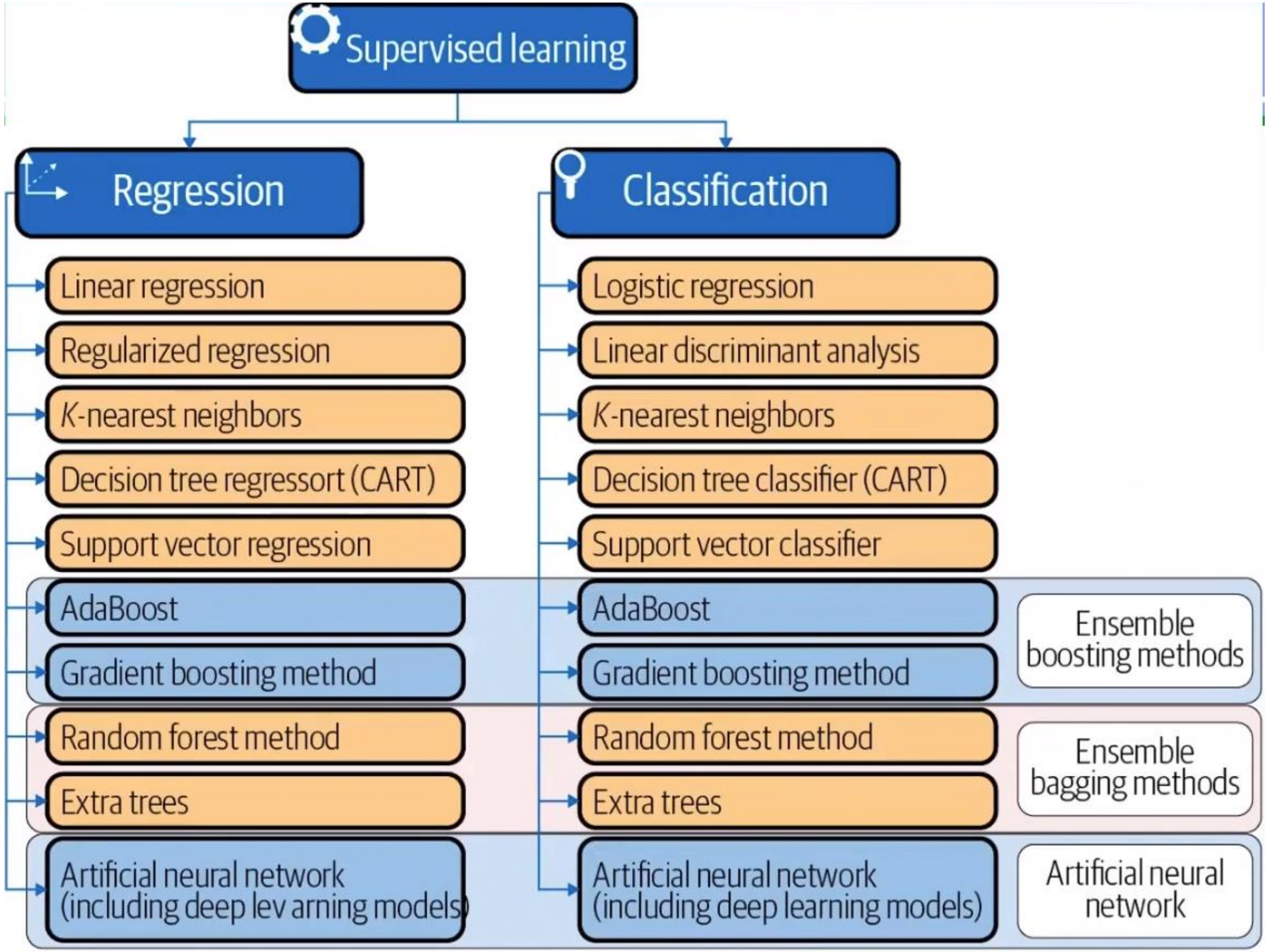
TECHPRO EDUCATION

techproeducation.com

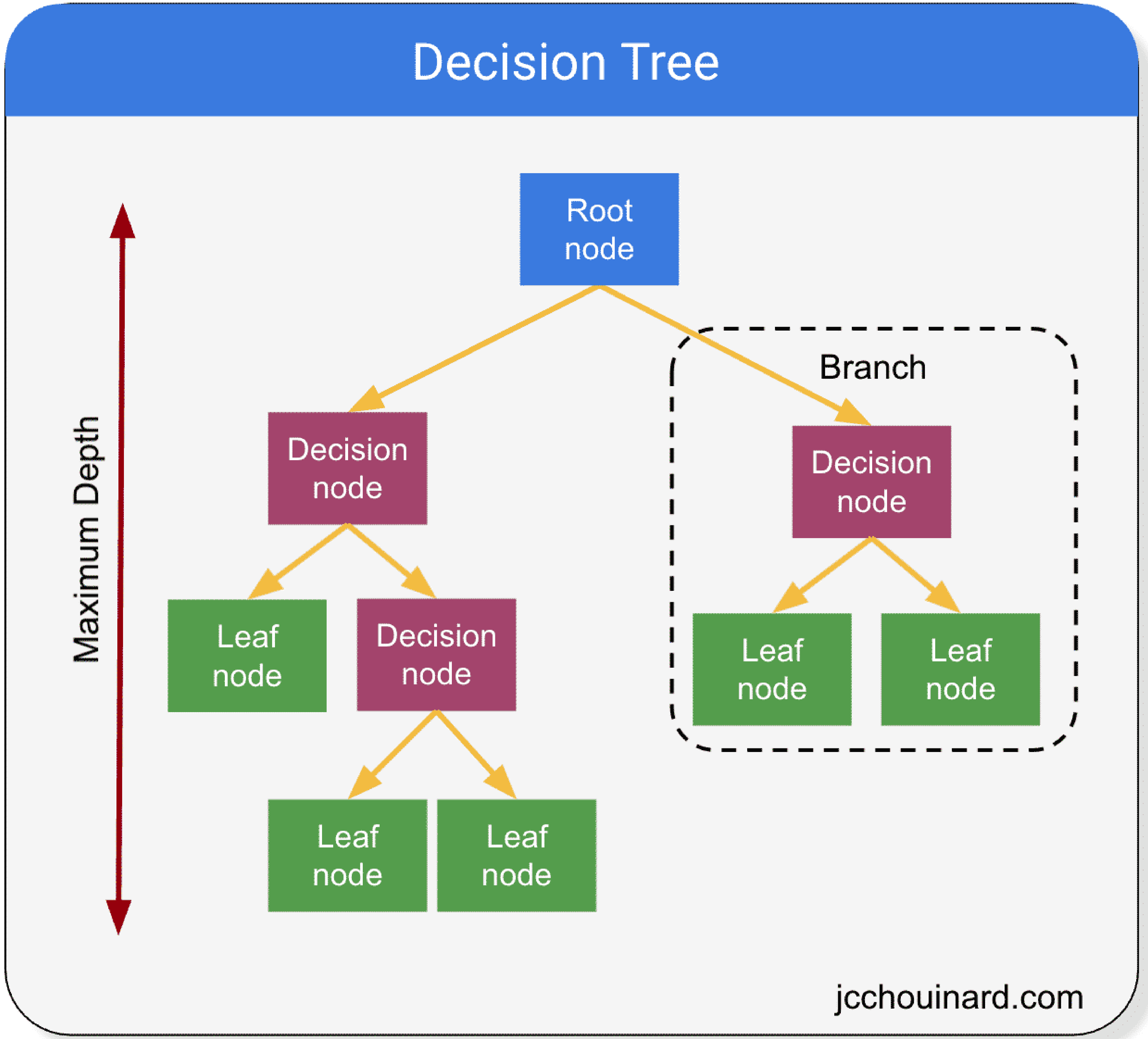
+1 (585) 304 29 59

DECISION TREE

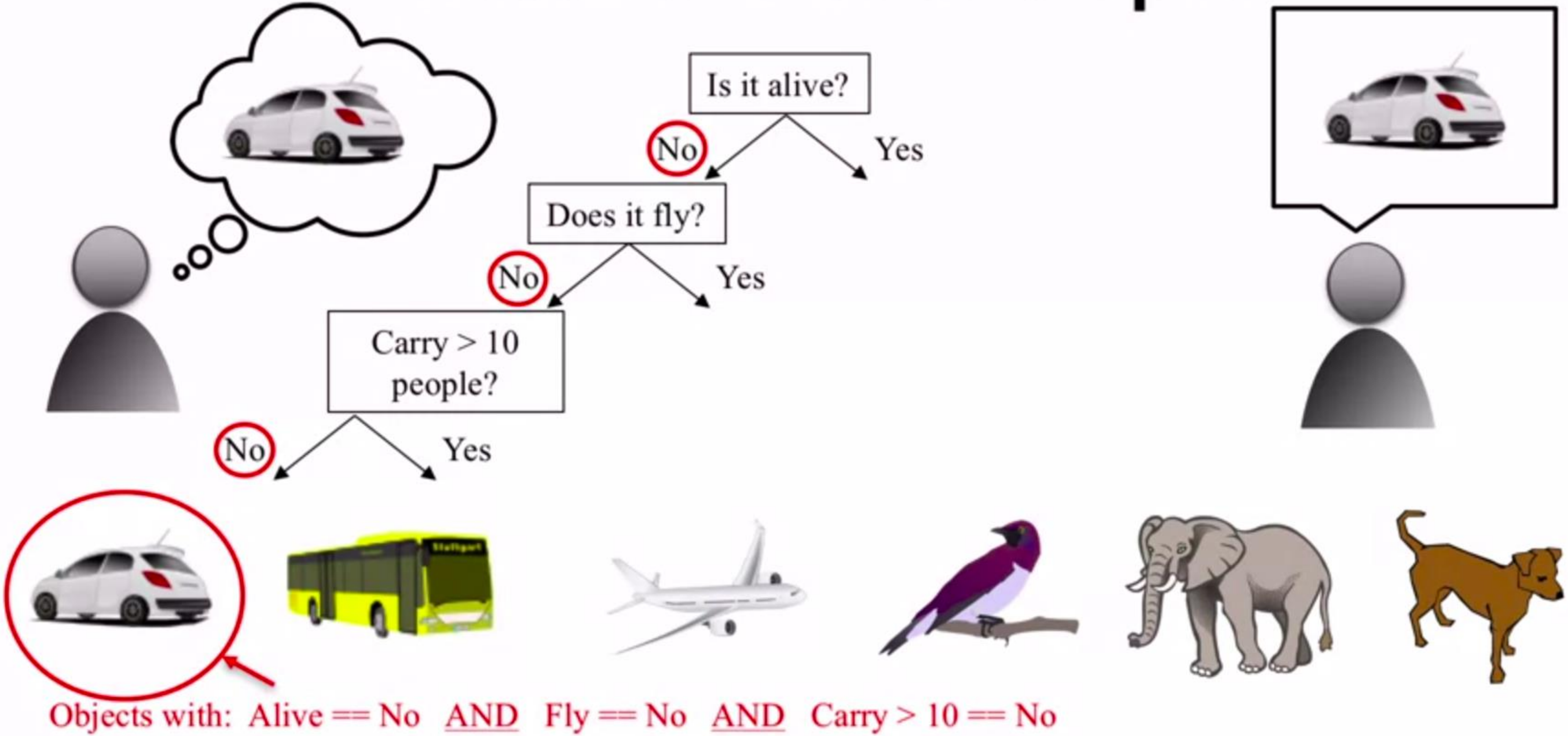
Where We Are?

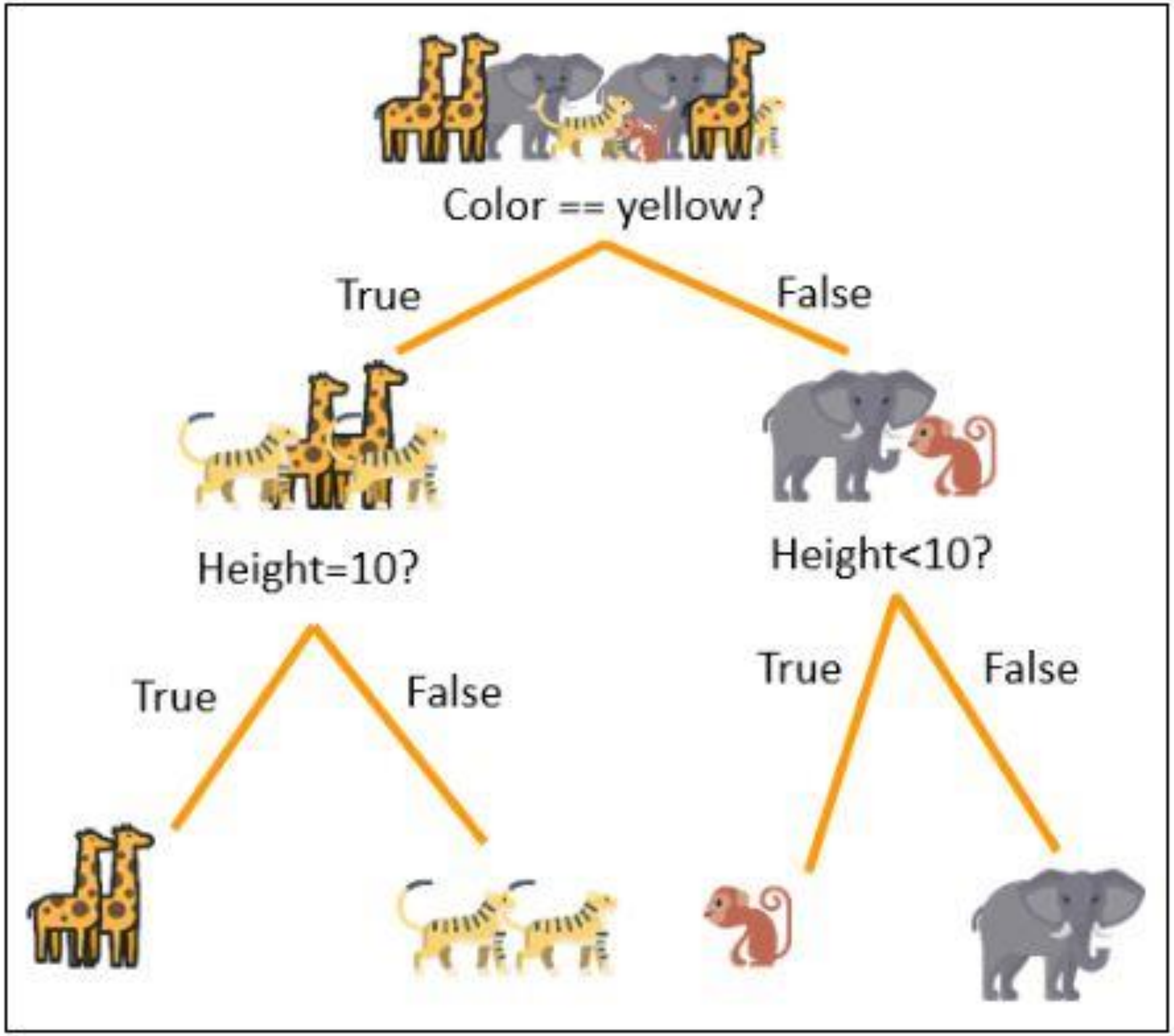

Supervised learning
Regression
Classification
Linear regression
Regularized regression
K-nearest neighbors
Decision tree regressort (CART)
Support vector regression
AdaBoost
Gradient boosting method
Random forest method
Extra trees
Artificial neural network
(including deep lev arning models)
Logistic regression
Linear discriminant analysis
K-nearest neighbors
Decision tree classifier (CART)
Support vector classifier
AdaBoost
Gradient boosting method
Random forest method
Extra trees
Artificial neural network
(including deep learning models)
Ensemble boosting methods
Ensemble bagging methods
Artificial neural network

Decision Tree

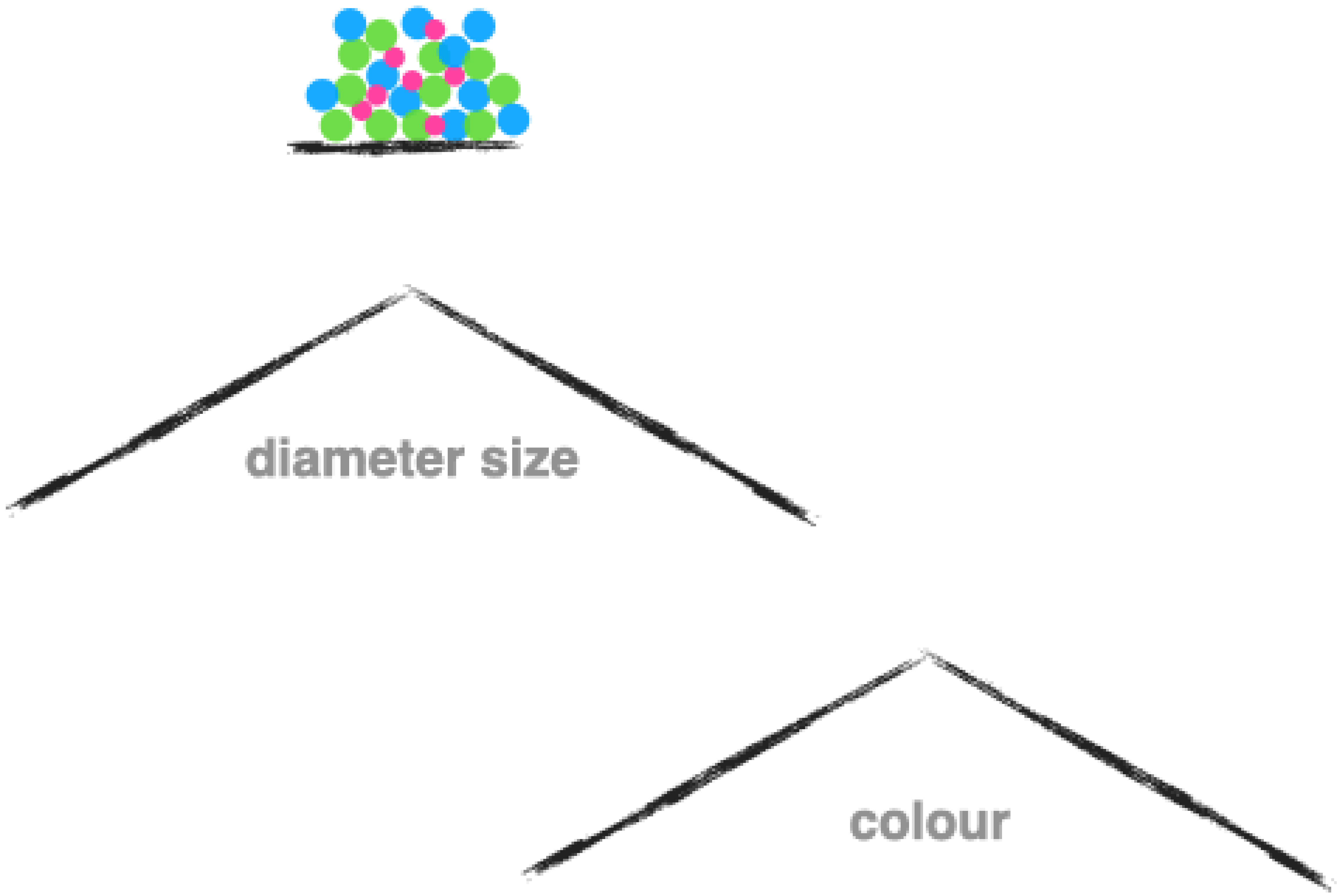


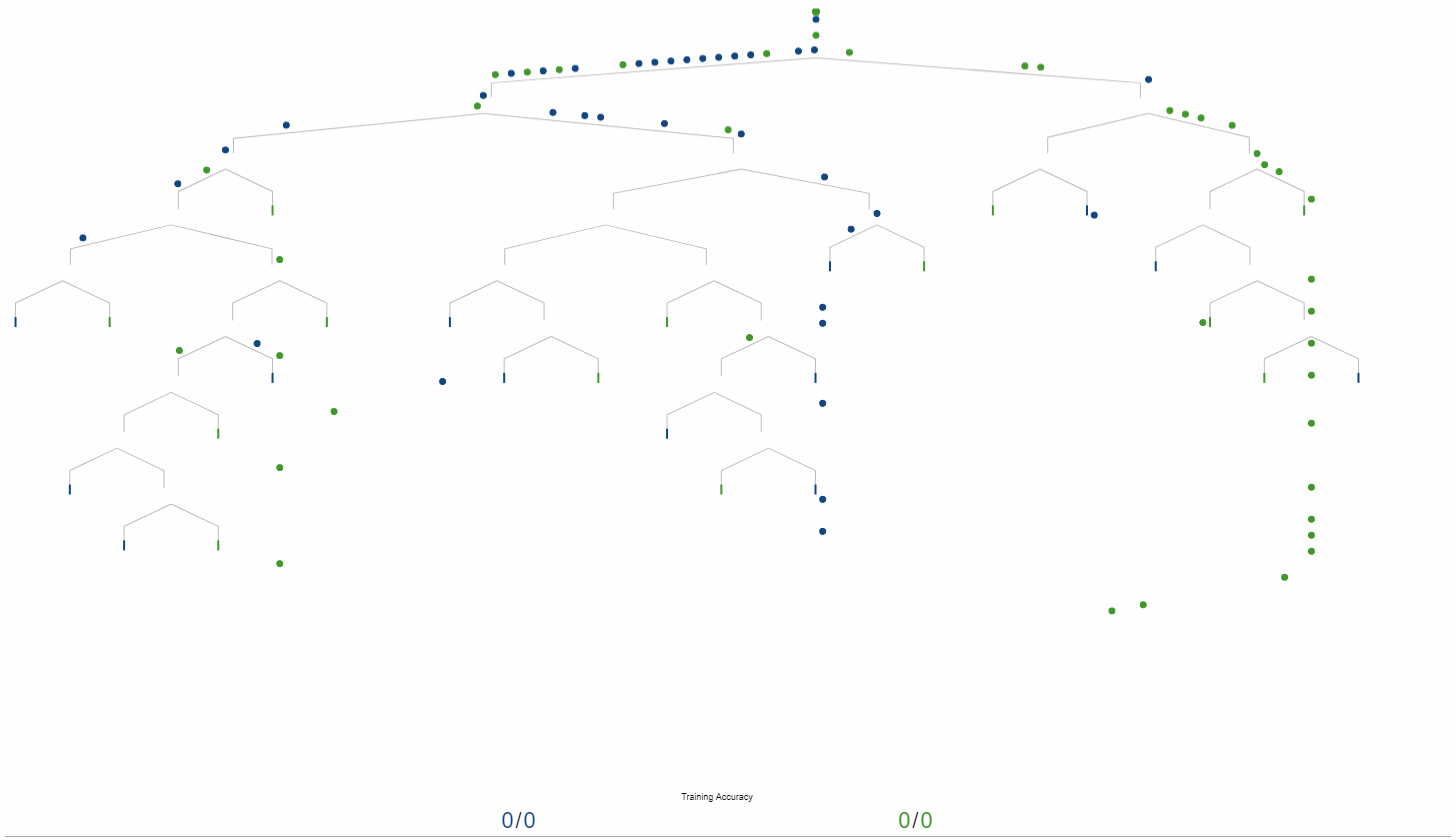
Decision Tree Example



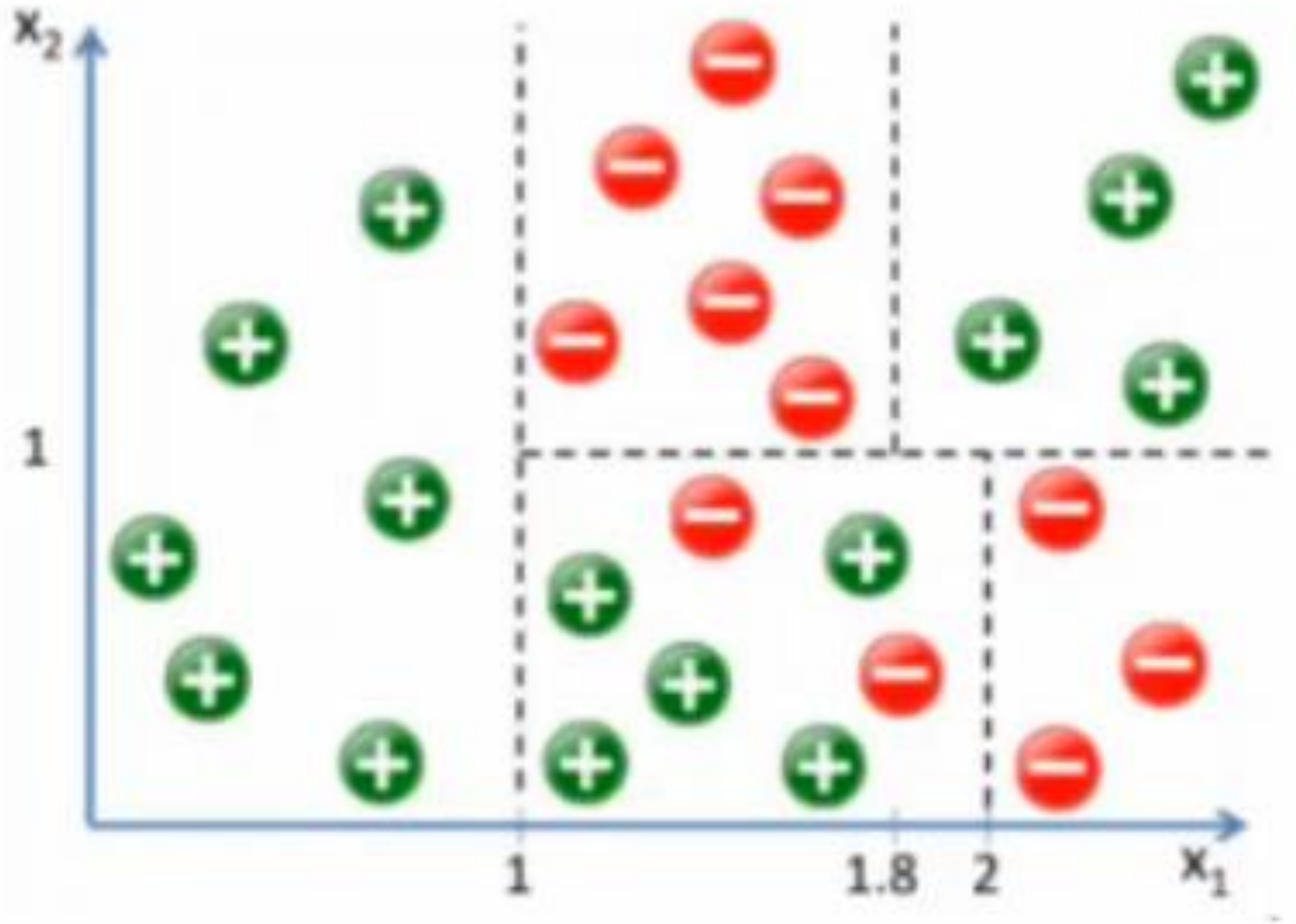


TECHPRO EDUCATION


diameter size
colour



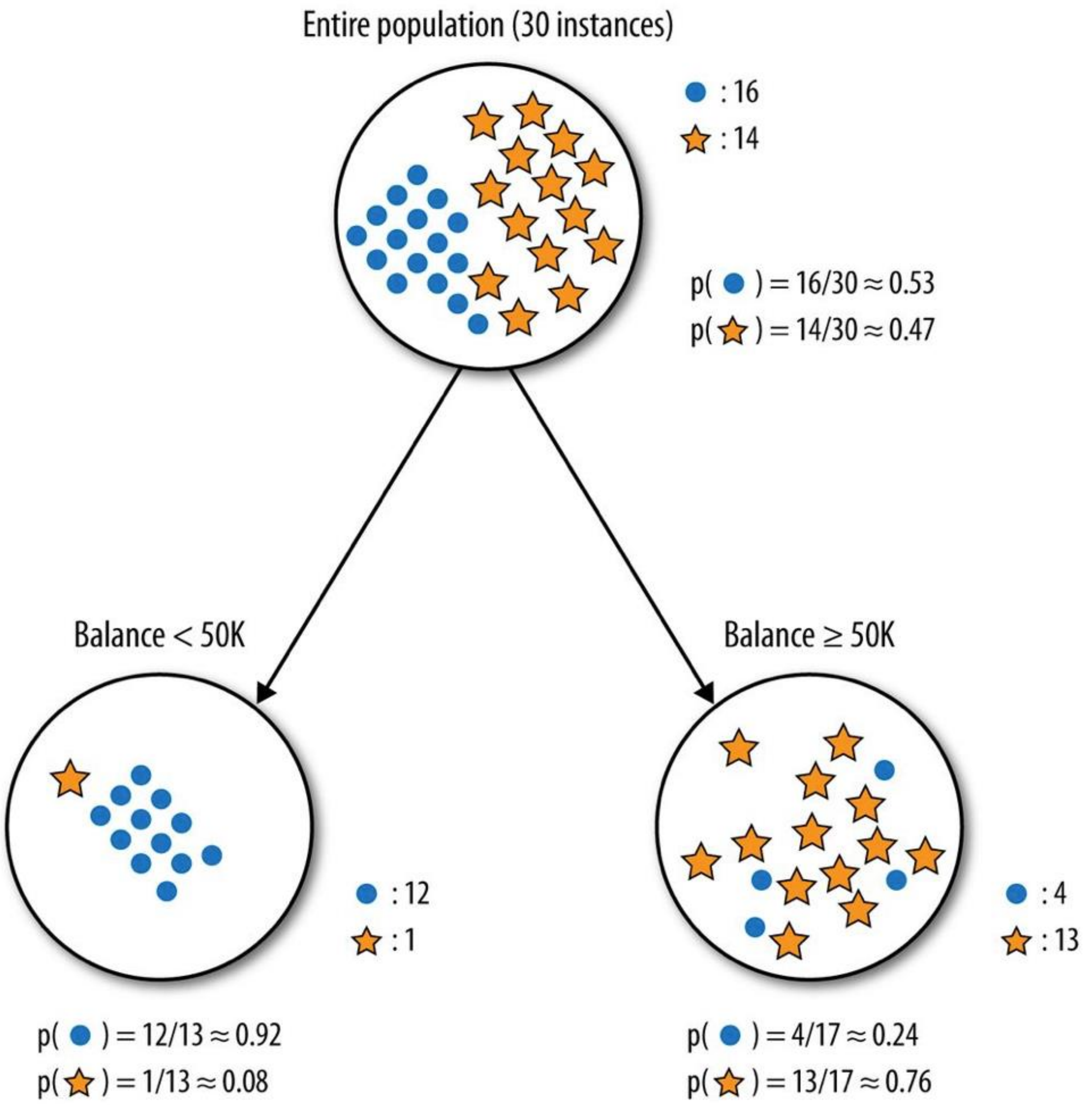
Training Accuracy
0/0
0/0



x_2
1
1
1.8
2
x_1

TEST DATA

DECISION TREE

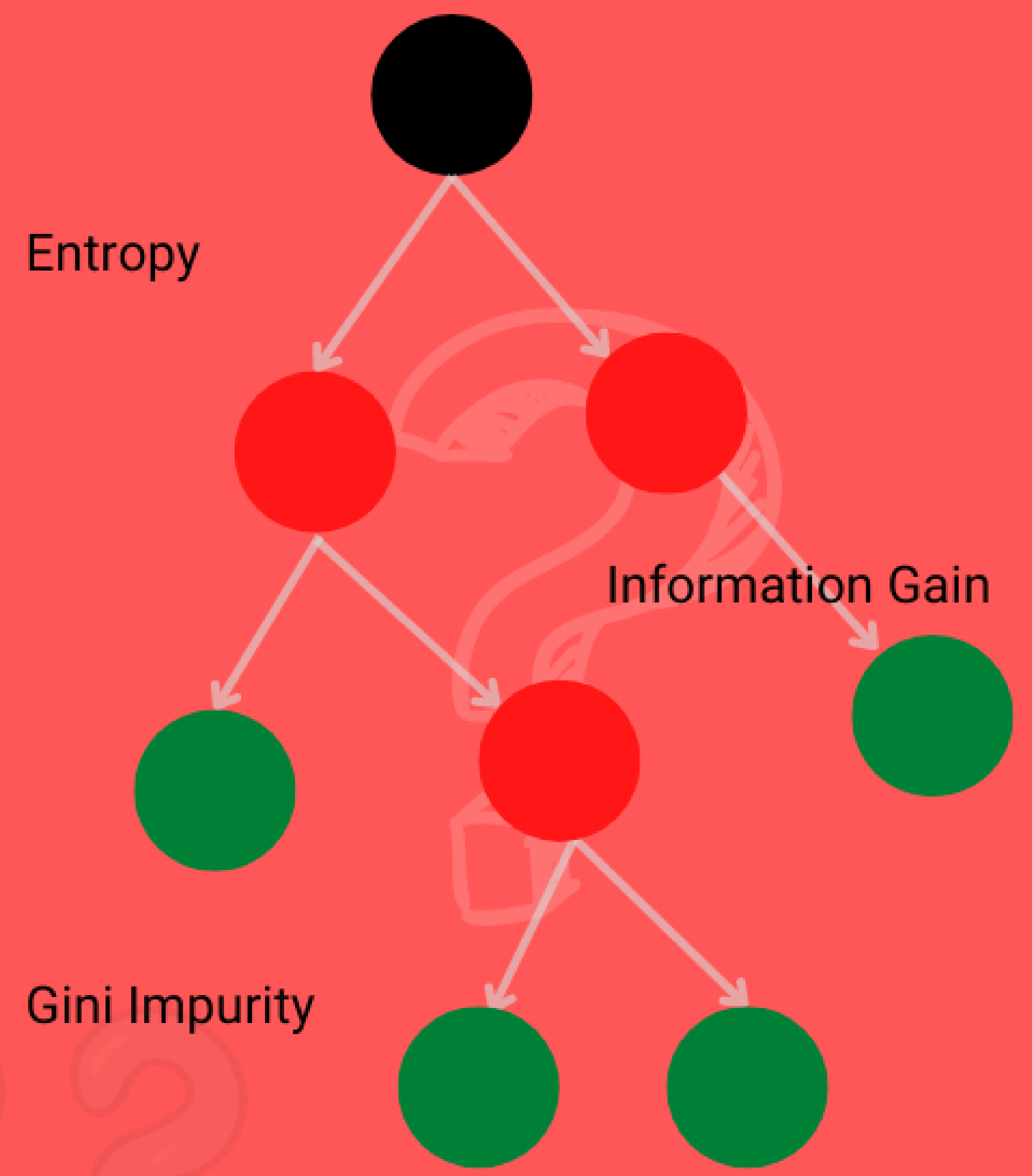


TECHPRO EDUCATION

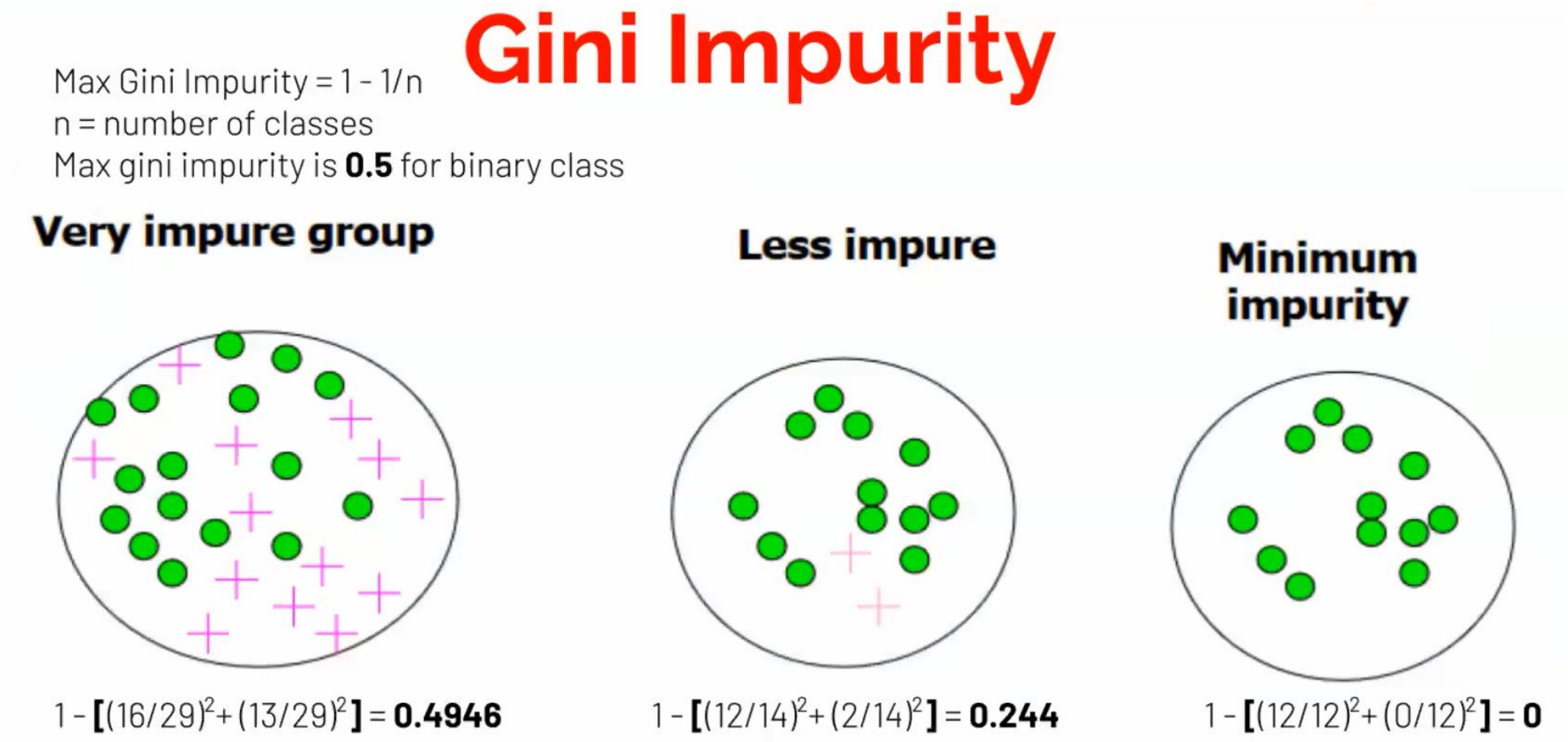
Decision Tree Splitting criteria

ID3

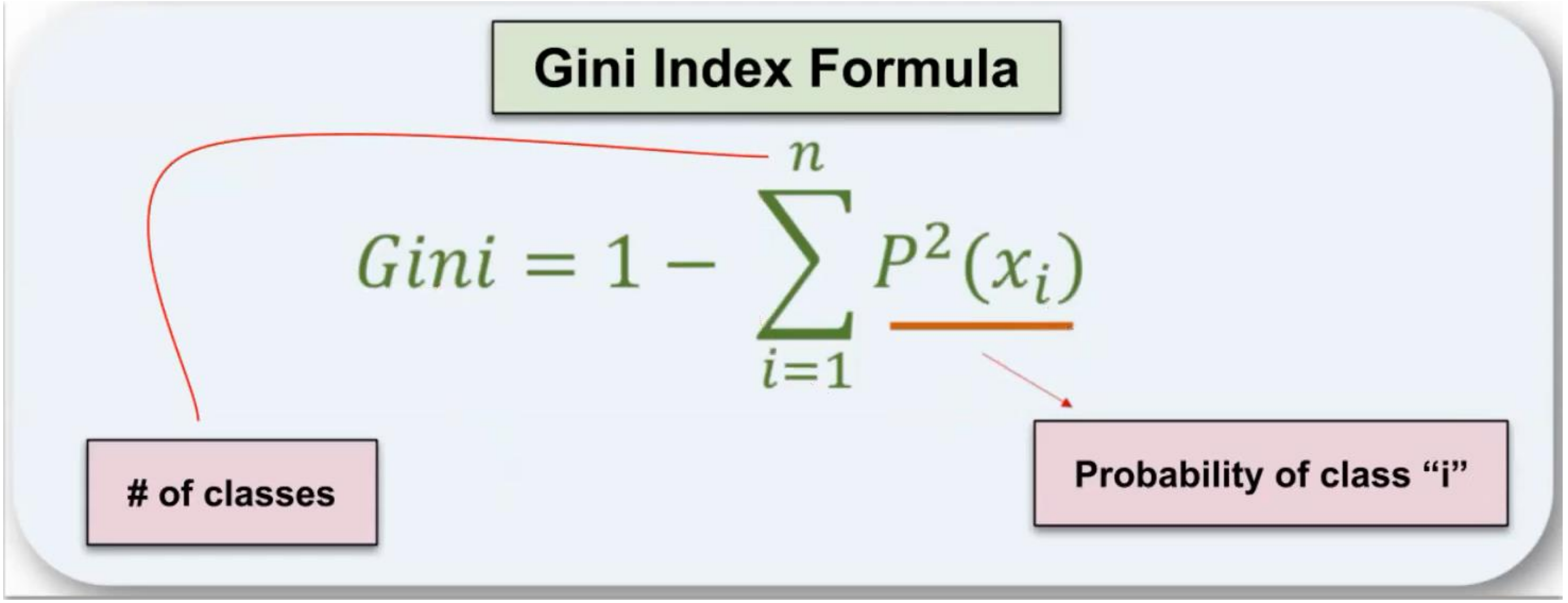
CART



Gini Impurity



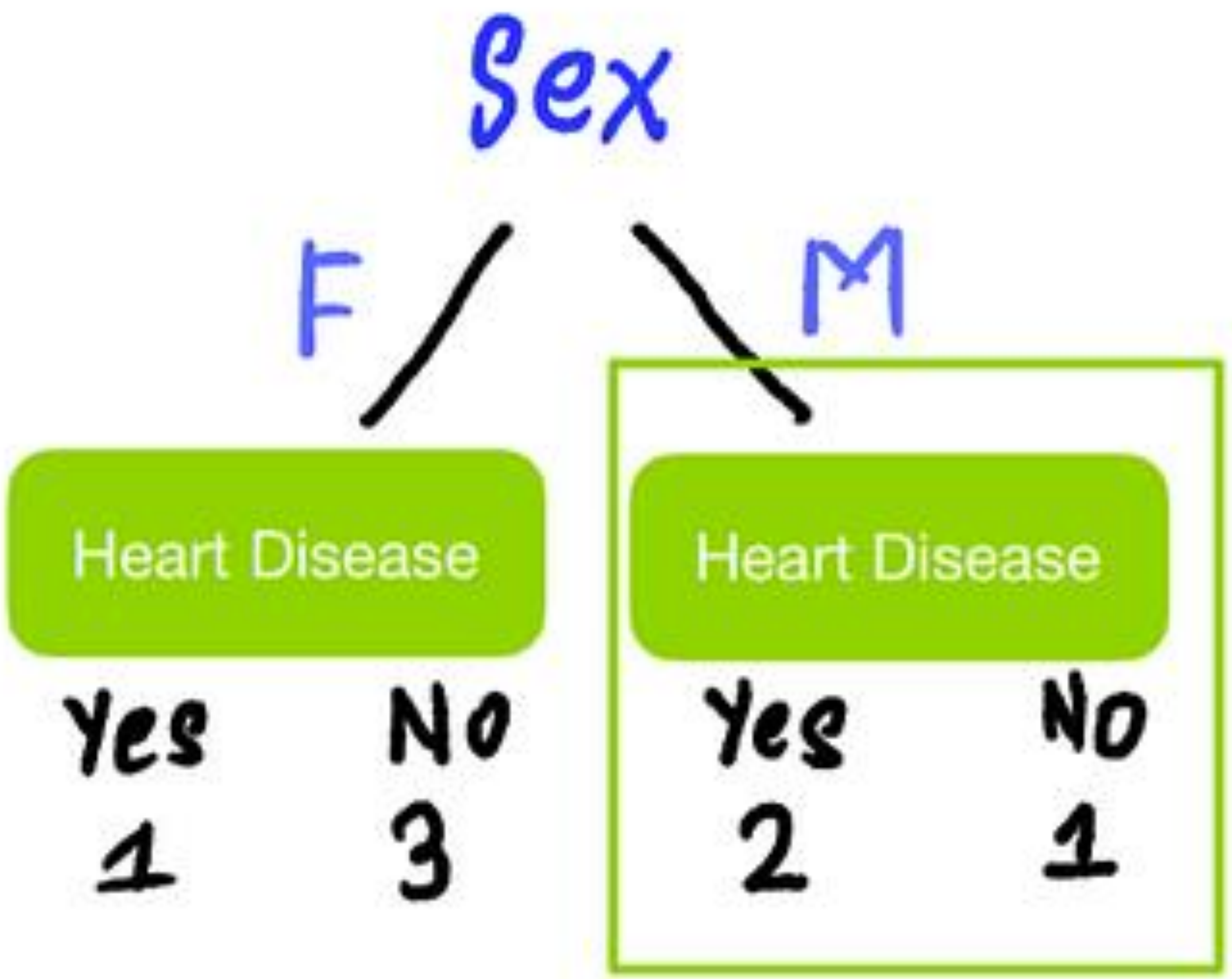
Gini Impurity



Another commonly used formula is:

$$Gini\ impurity = 1 - \Sigma\left(p(i) * \left(1 - p(i)\right)\right)$$

Gini Impurity



$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Gini Impurity of the right leaf

$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2$$

$$= 0.444$$

TECHPRO EDUCATION

Gini Impurity

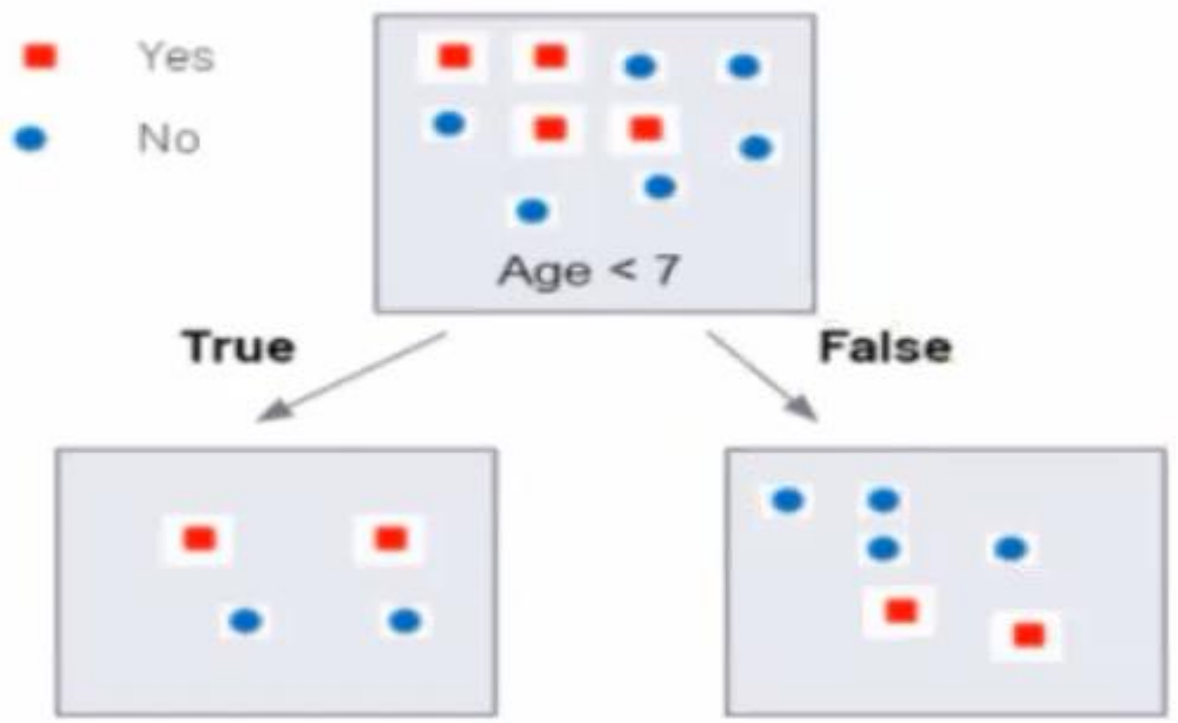
Age	Ability	Footballer (tiny stars)
10	Yes	No
8	No	No
6	Yes	Yes
6	No	No
8	Yes	Yes
8	Yes	Yes
6	Yes	Yes
10	Yes	No
10	No	No
6	No	No

Selection Feature For Root Node With Gini Impurity

Age	Ability	Footballer (tiny stars)
10	1	No
8	0	No
6	1	Yes
6	0	No
8	1	Yes
8	1	Yes
6	1	Yes
10	1	No
10	0	No
6	0	No

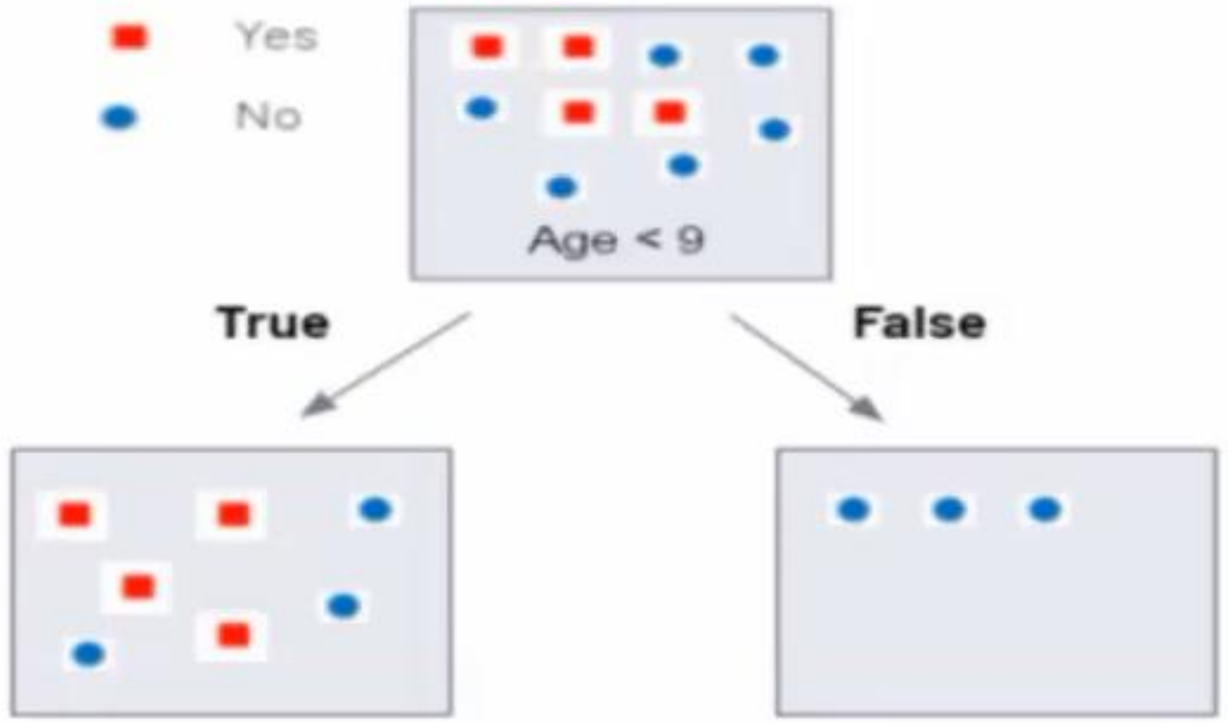
Sorting Age : 6 - 8 - 10
Sorting Ability : 0 - 1

For Age Feature



$1 - [(2/4)^2 + (2/4)^2] = \mathbf{0.5}$ $\quad$ $1 - [(4/6)^2 + (2/6)^2] = \mathbf{0.44}$

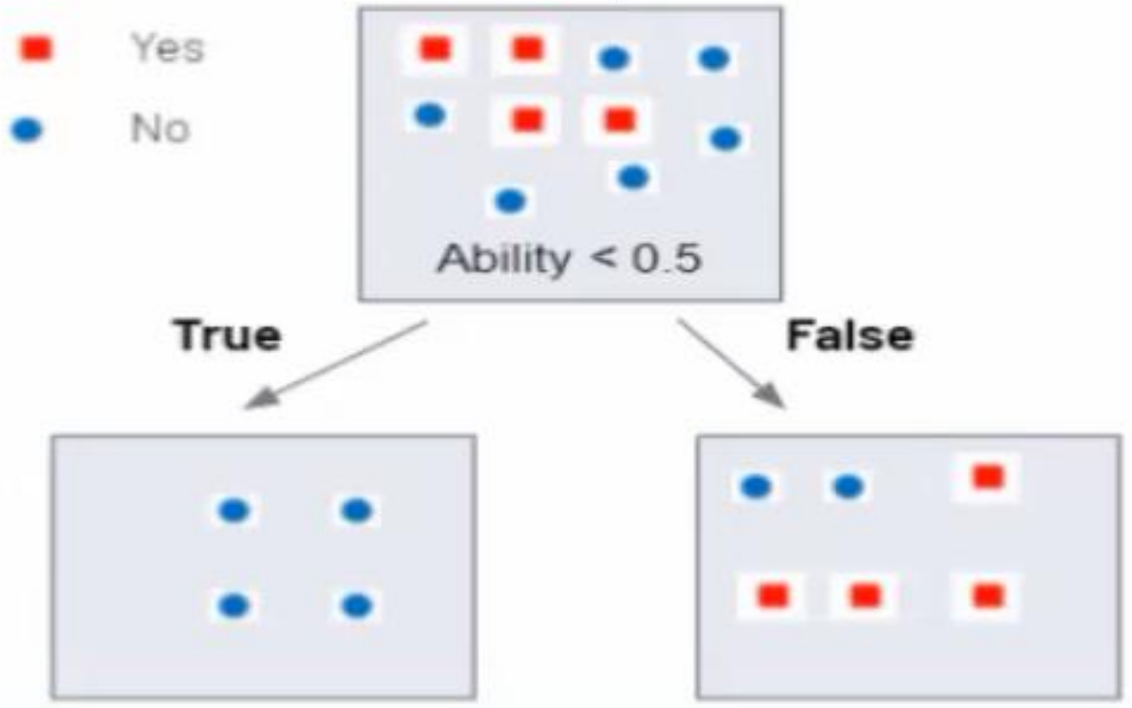
$4/10 \times \mathbf{0.5} + 6/10 \times \mathbf{0.44} = \mathbf{0.464}$



$1 - [(4/7)^2 + (3/4)^2] = \mathbf{0.49}$ $\quad$ $1 - [(0/3)^2 + (3/3)^2] = \mathbf{0}$

$7/10 \times \mathbf{0.49} + 3/10 \times \mathbf{0} = \mathbf{0.34}$

For Ability Feature

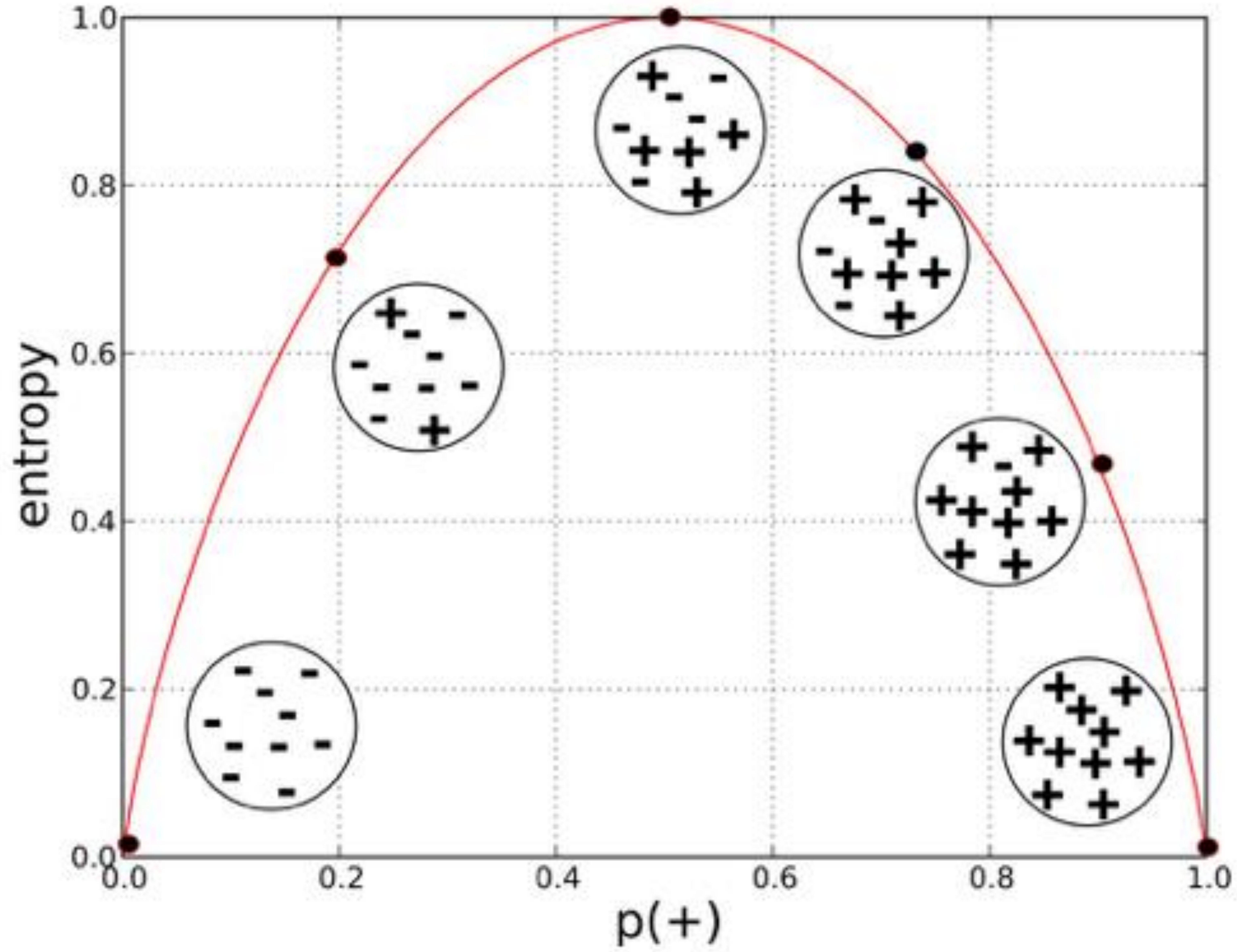


$1 - [(0/4)^2 + (4/4)^2] = \mathbf{0}$ $\quad$ $1 - [(4/6)^2 + (2/6)^2] = \mathbf{0.44}$

$4/10 \times \mathbf{0} + 6/10 \times \mathbf{0.44} = \mathbf{0.26}$

Root Node

Entropy



Entropy

The formula for Entropy is shown below:

$$E(S) = -p_{(+)} \log p_{(+)} - p_{(-)} \log p_{(-)}$$

Here,

- p_+ is the probability of positive class
- p_- is the probability of negative class
- S is the subset of the training example

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

Entropy

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

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

Entropy

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

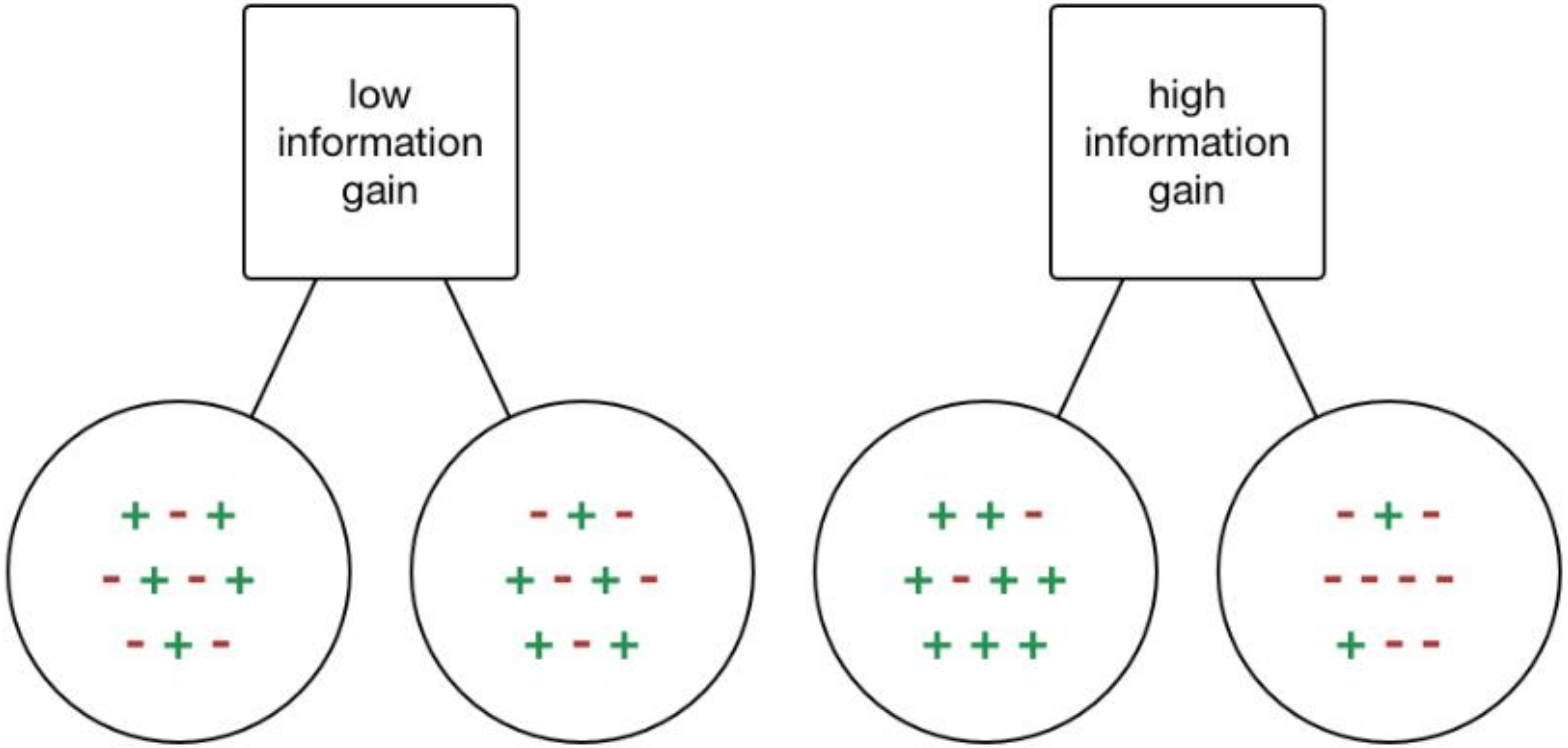
		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

Information Gain

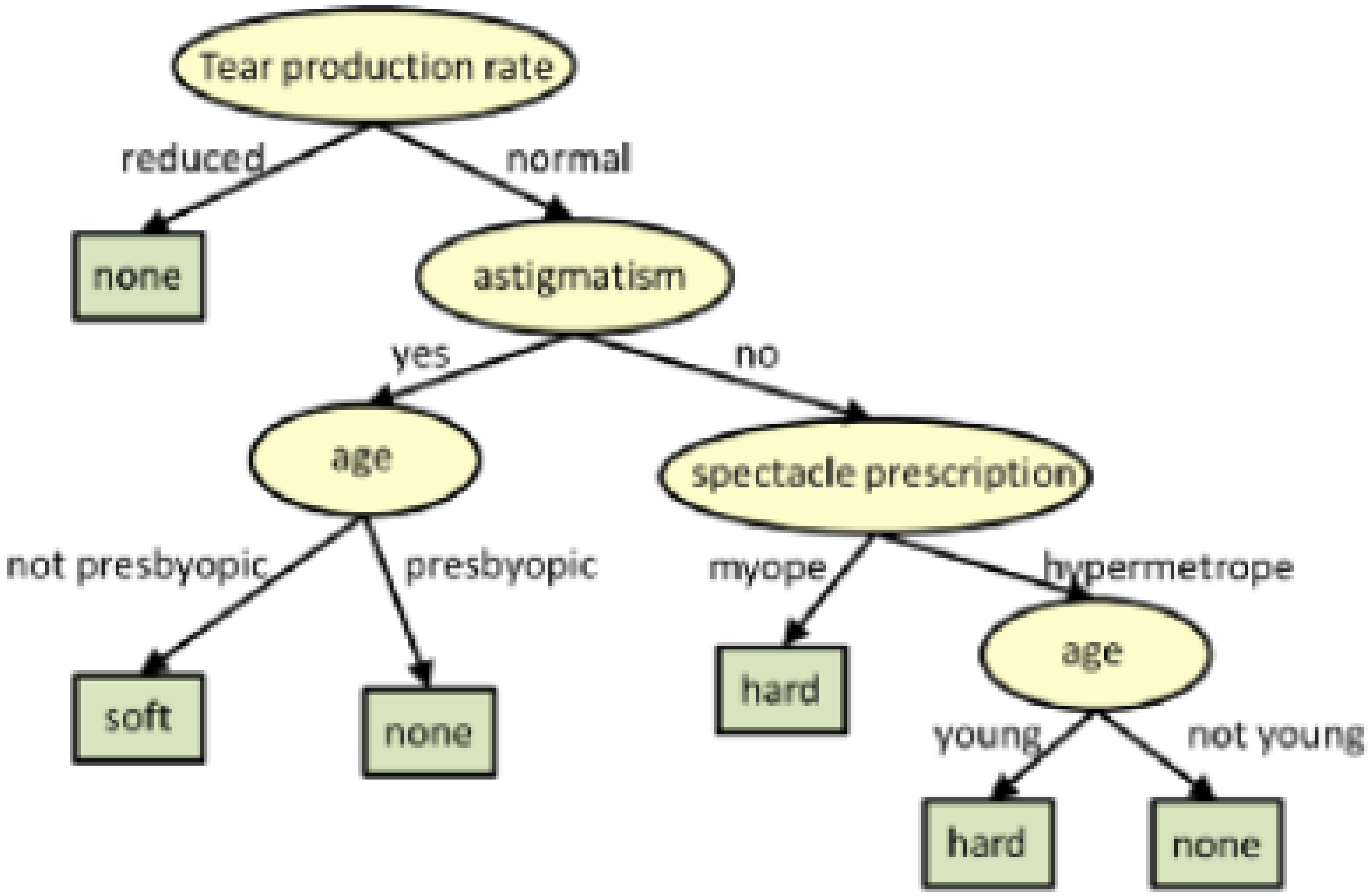


Information Gain

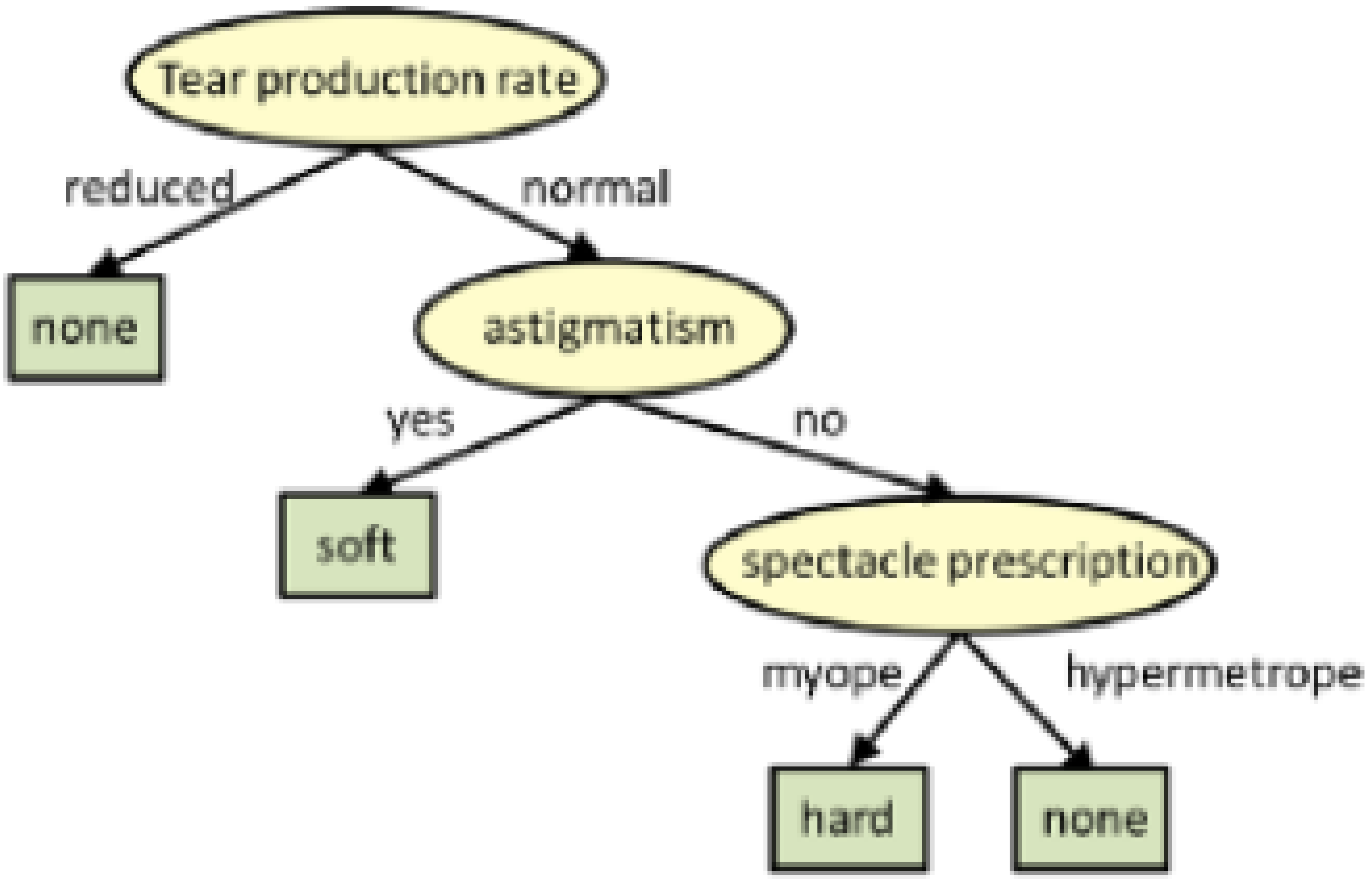
Information Gain(T,X) = Entropy(T) - Entropy(T, X)

IG(PlayGolf, Outlook) = E(PlayGolf) - E(PlayGolf, Outlook)
= 0.940 - 0.693
= 0.247

Pruned Tree



Original Tree



Pruned Tree

Karar ağacının nasıl bölüneceğini nasıl belirleriz?

- Karar ağaçlarında bölünme (veya dal ayrımı), veri setindeki homojenliği maksimize etmek için yapılır.

- Yani, bir ağacın her bir dalında, sonucun mümkün olduğunca bir sınıfa özgü olması hedeflenir.

- Bu, bilgi kazancı (Information Gain), Gini saflığı (Gini Impurity), veya entropi gibi ölçümler kullanılarak belirlenir.

Tea break...

10:00

Start Stop Reset mins: 10 secs: 0 type: