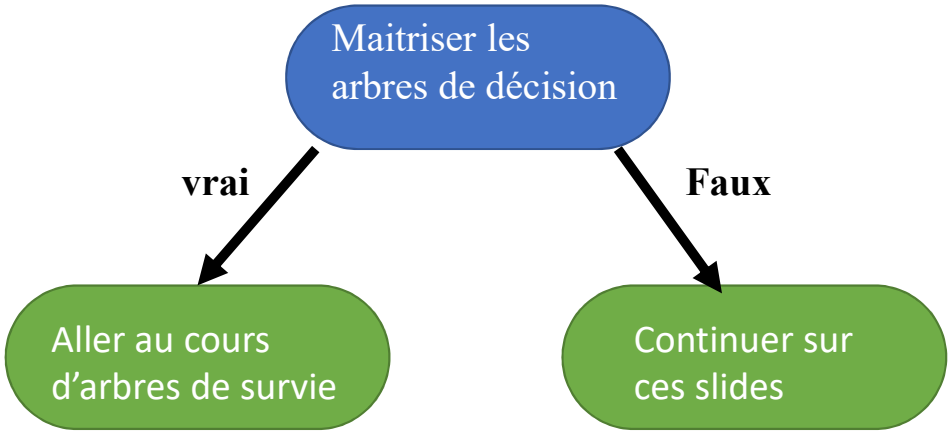
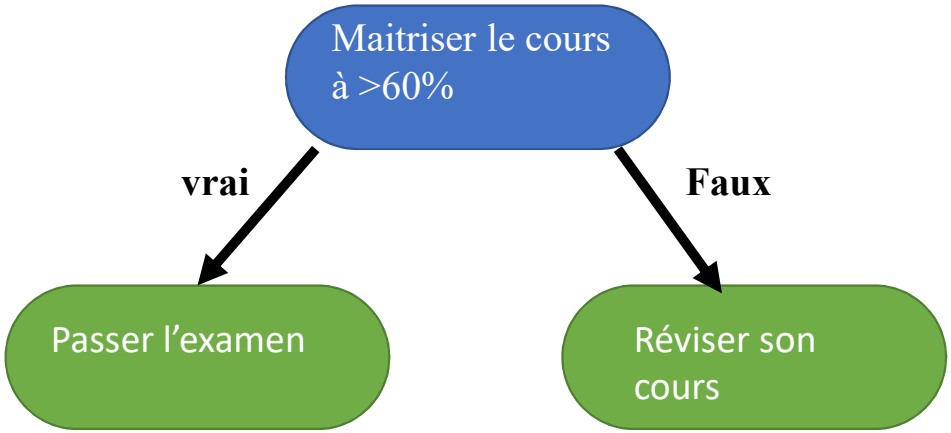
Analyse de Survie
Professeur Abdellatif El Afia

Arbres de Décision

Arbres de décision

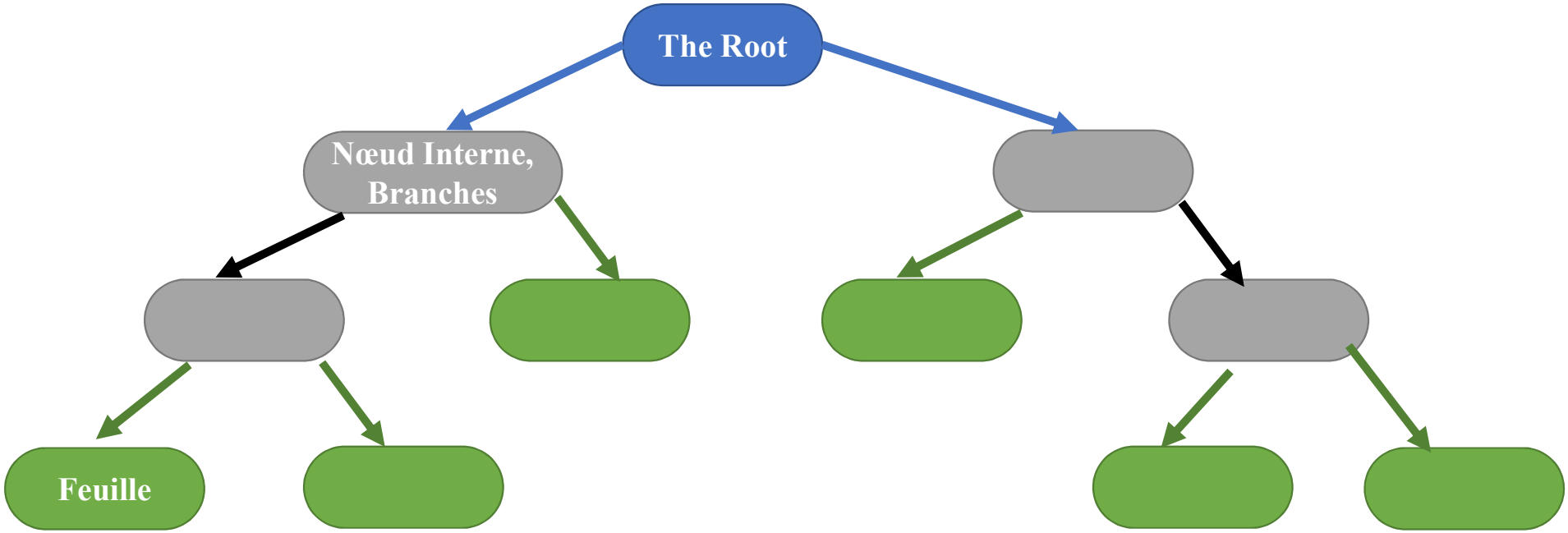


Arbre de classification



Arbre de régression

Arbres de décision

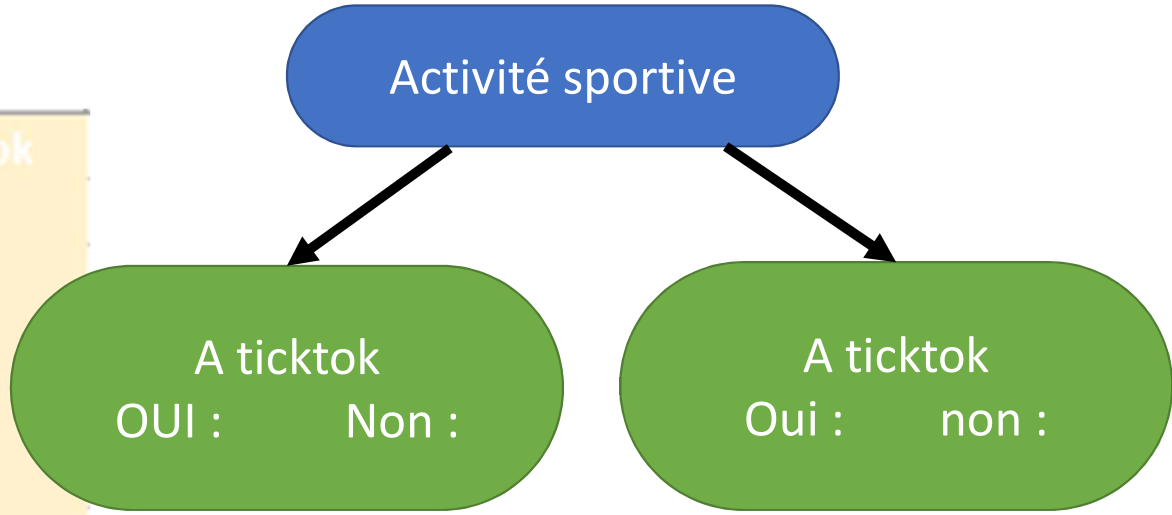


Arbres de décision

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non

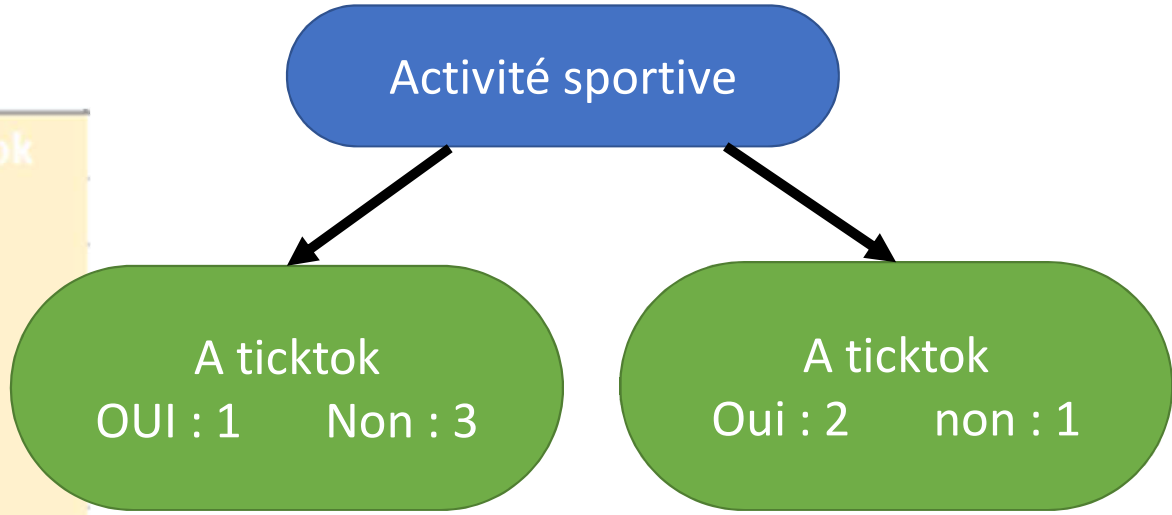
Arbres de décision

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non



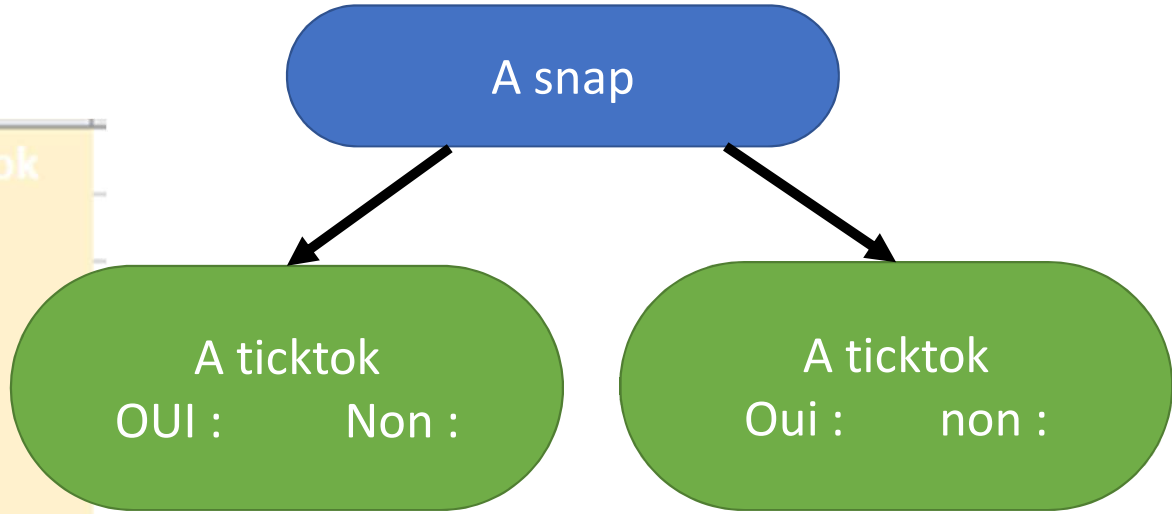
Arbres de décision

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non



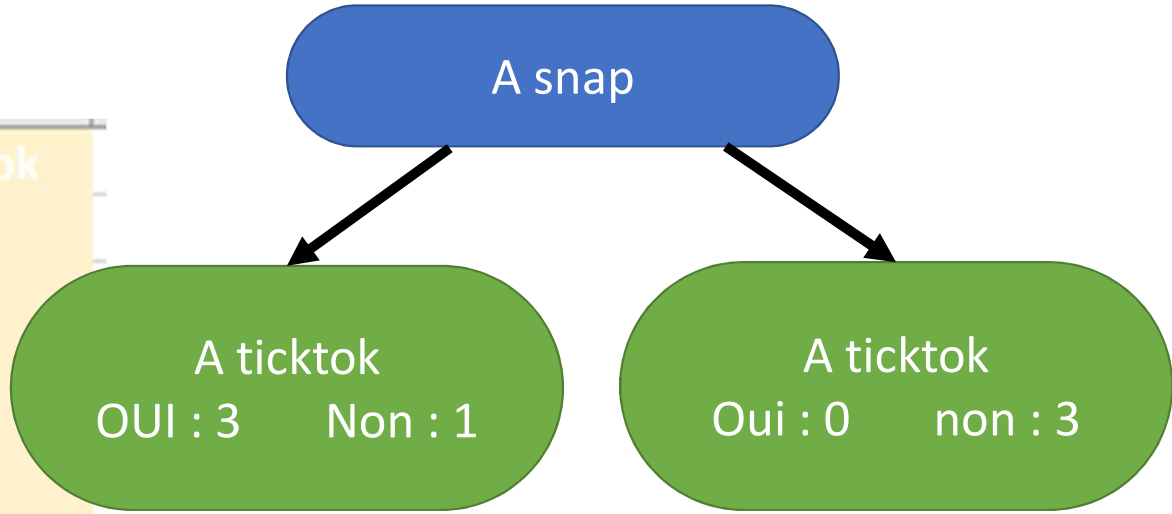
Arbres de décision

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non

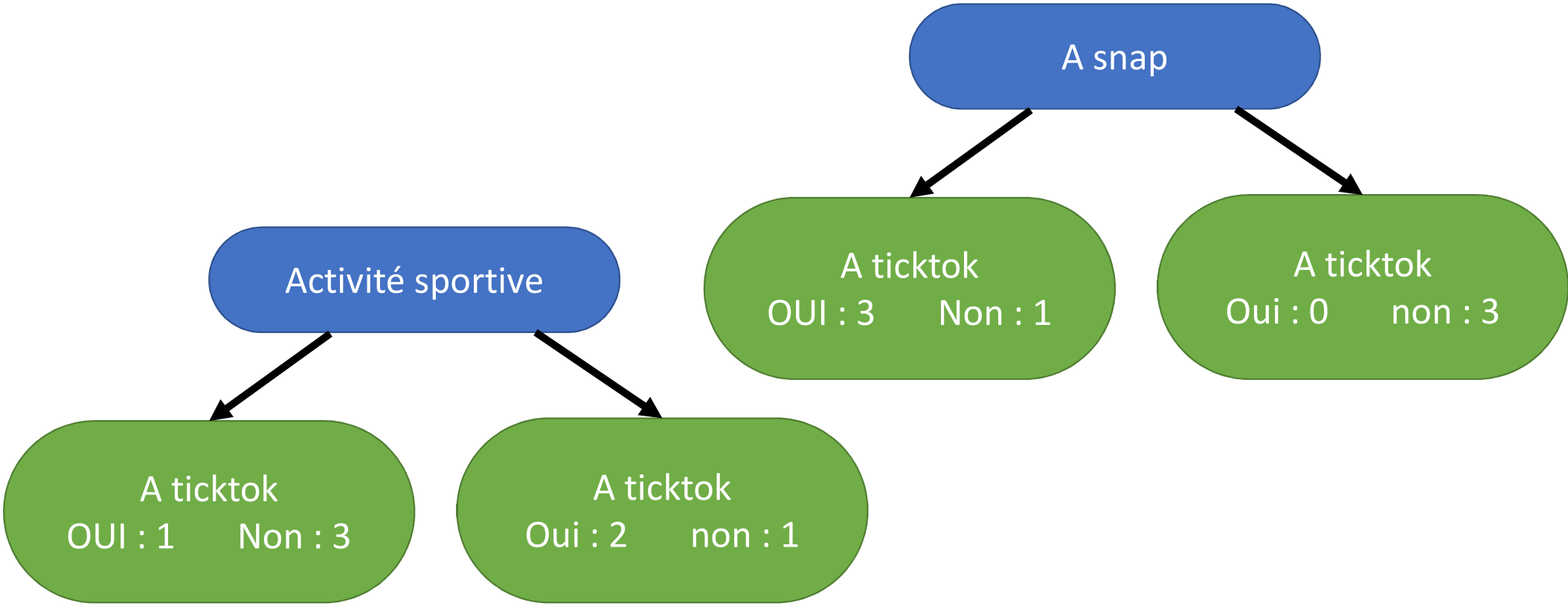


Arbres de décision

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non



Arbres de décision



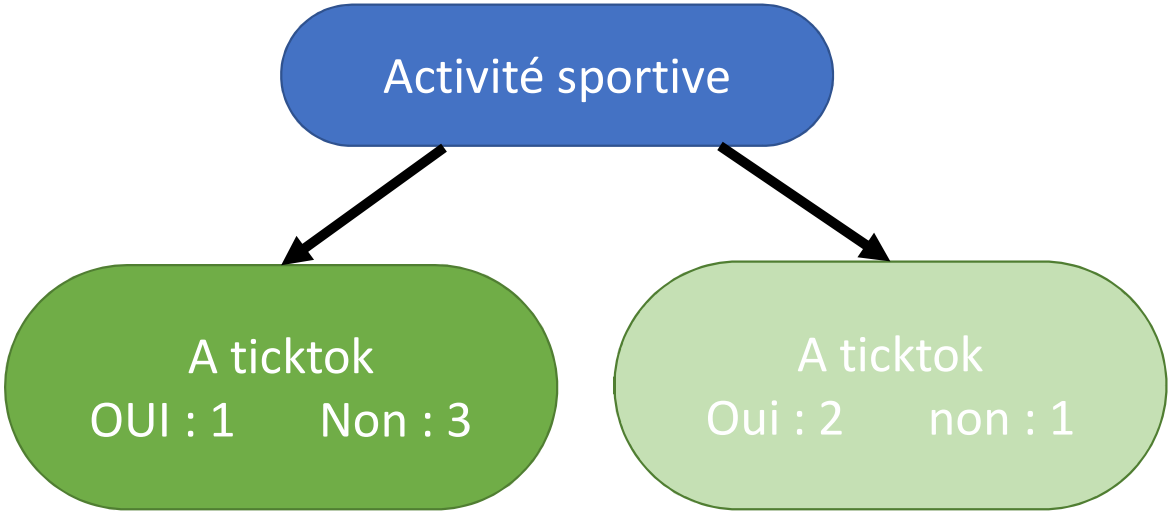
Gini Impurity

Gini impurity for a leaf = $1 - P(oui)^2 - P(non)^2$

Total Gini Impurity = Somme of weighted Gini imprities

Gini Impurity

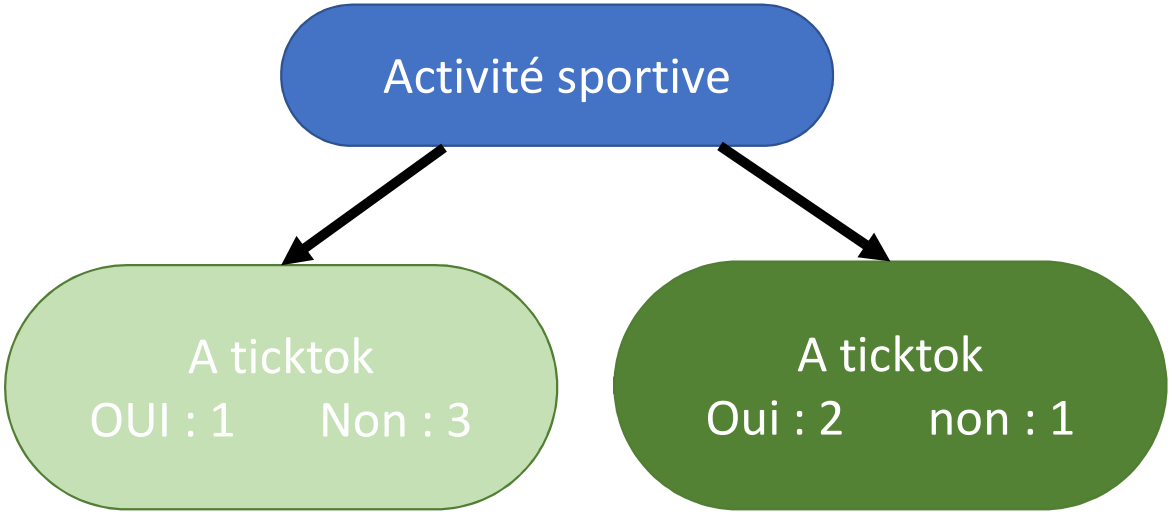
Gini impurity for a leaf = 1 – $P(oui)^2$ - $P(non)^2$



Gini Impurity = 1 – $(1/4)^2$ - $(3/4)^2$
= 0.375

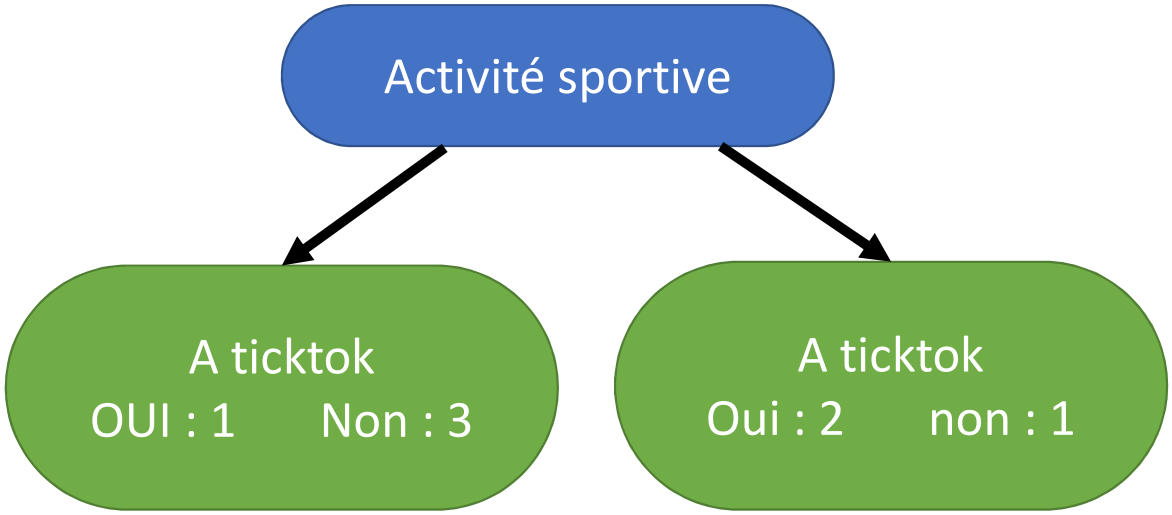
Gini Impurity

Gini impurity for a leaf = 1 – $P(oui)^2$ - $P(non)^2$



Gini Impurity = 1 – $(2/3)^2$ - $(1/3)^2$
= 0.444

Gini Impurity

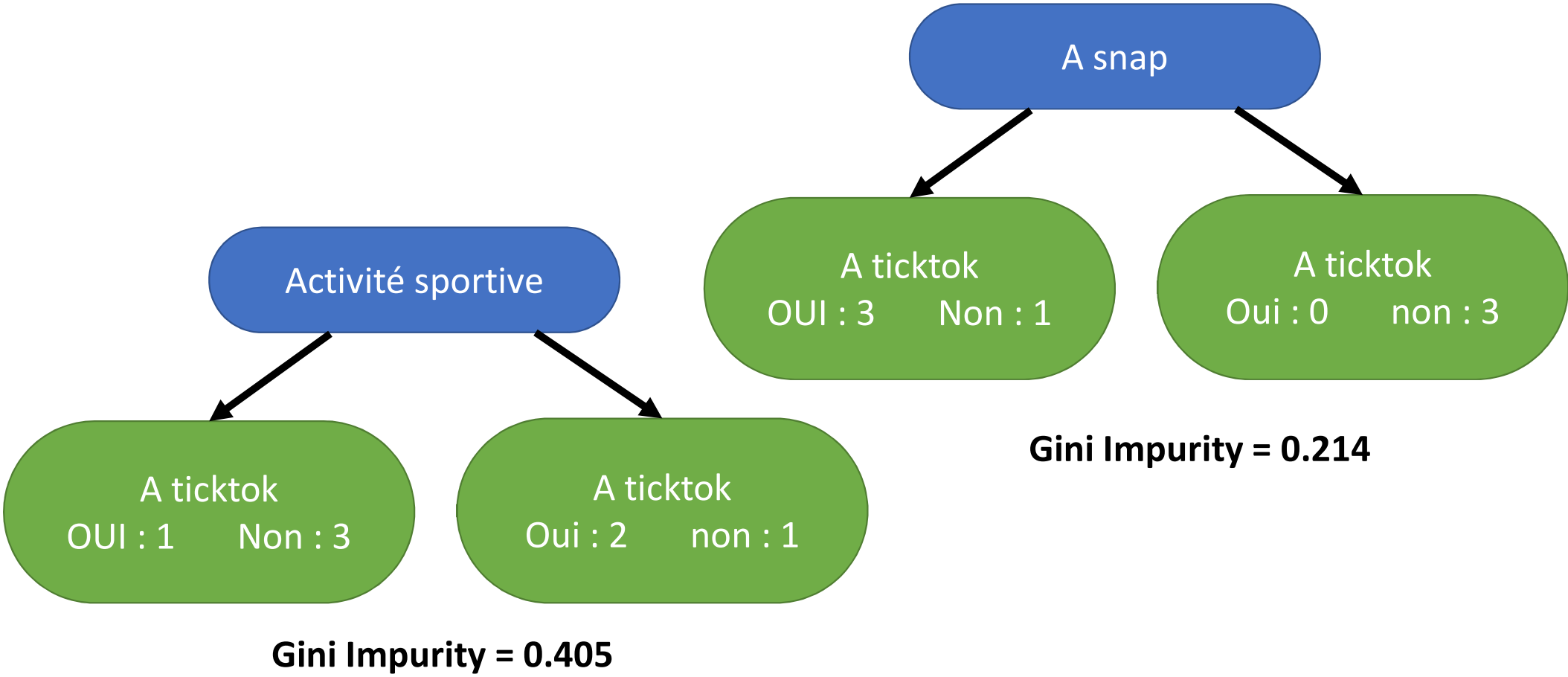


Gini Impuity	**0.375**	**0.444**

Total Gini Impurity = Somme of weighted Gini imprities

= 4/7 * 0.375 + 3/7 * 0.44 = 0.405

Gini Impurity



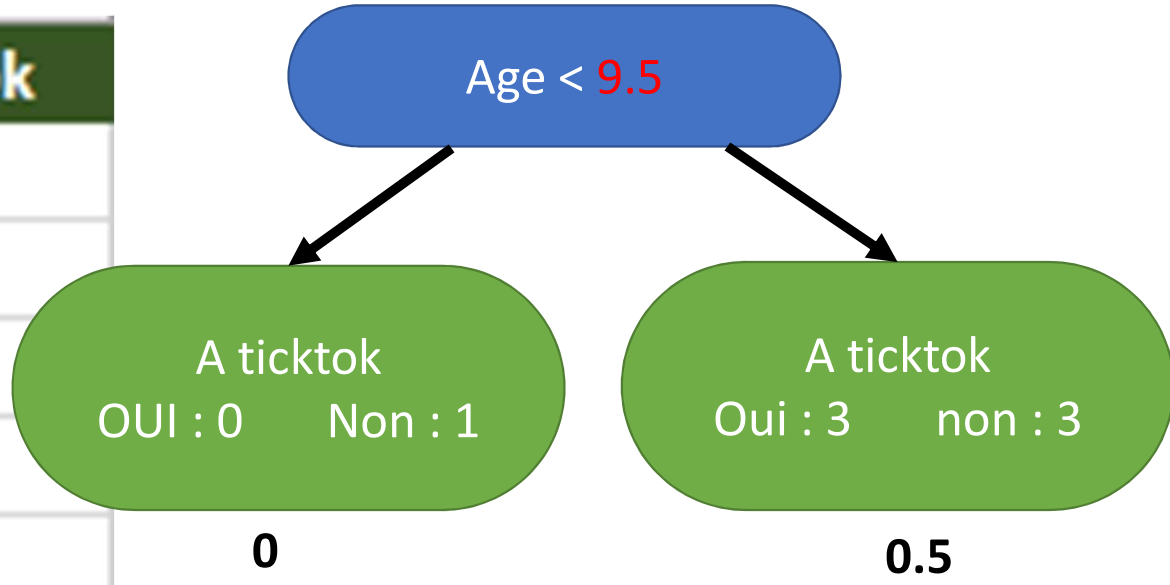
Gini Impurity For numerical data

	Age	A tioktok
	7	non
9.5		
	12	on
15		
	18	oui
26.5		
	35	oui
36.5		
	38	oui
44		
	50	non
66.5		
	83	non

- On commence par ordonner les valeurs dans un ordre croissant, puis on calcule la moyenne entre chaque deux valeurs successives.
- Chaque moyenne est un seuil qui partitionne la Data.
- On calcule le Gini Impurity pour chaque seuil et on choisi le seuil avec le Gini Impurity le plus petit.

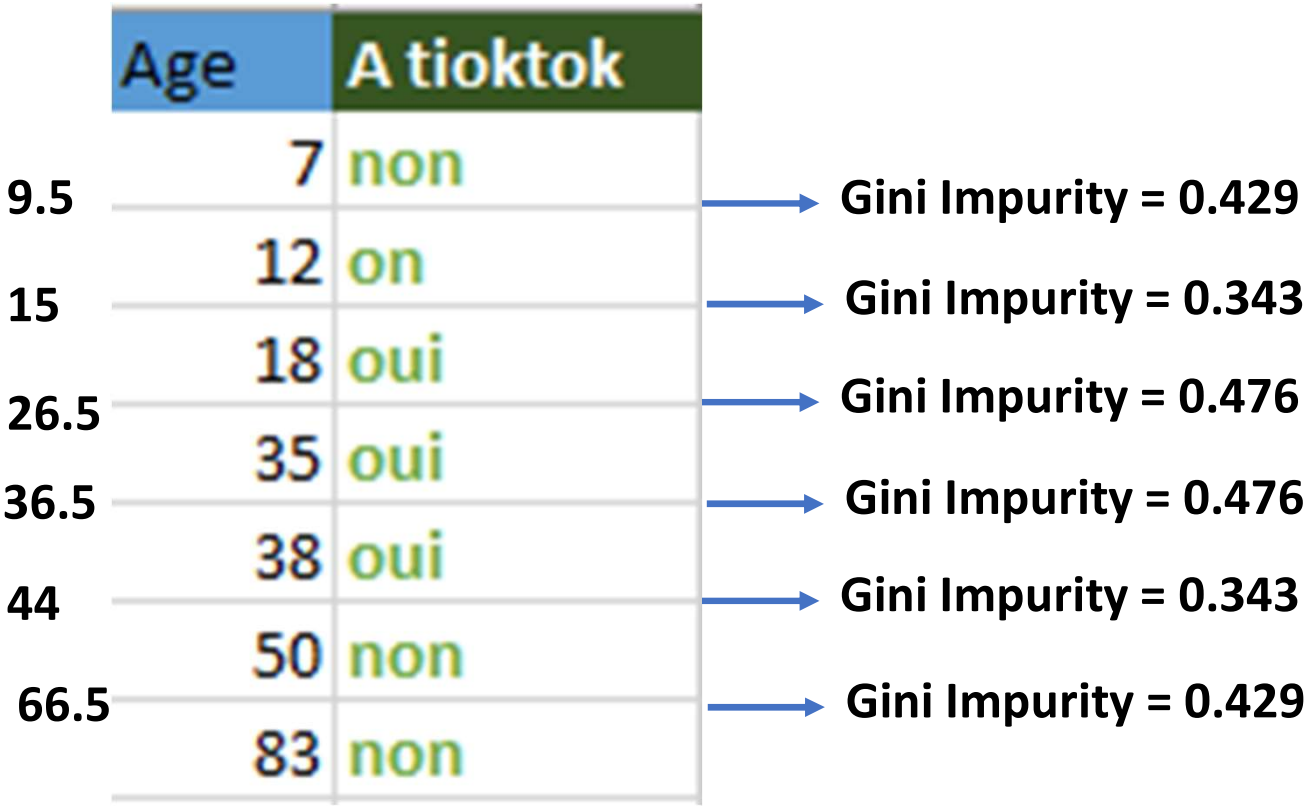
Gini Impurity For numerical data

	Age	A tioktok
	7	non
9.5		
	12	on
15		
	18	oui
26.5		
	35	oui
36.5		
	38	oui
44		
	50	non
66.5		
	83	non



Total Gini Impurity = 0.429

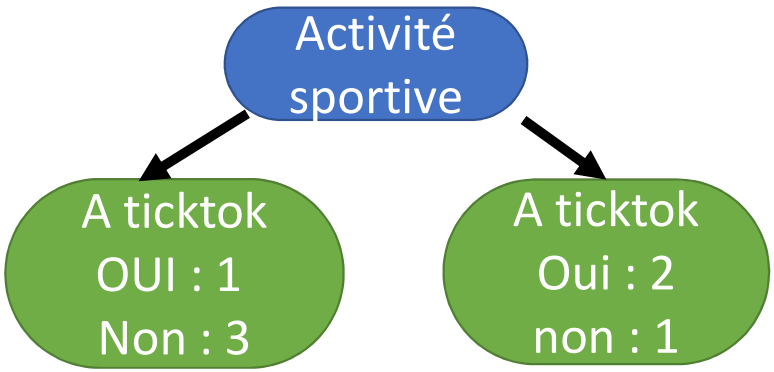
Gini Impurity For numerical data



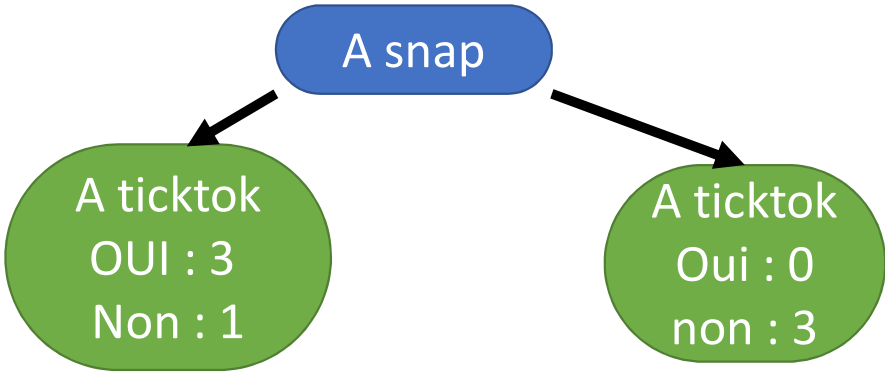
Gini Impurity For numerical data

	Age	A tioktok	
9.5	7	non	→ Gini Impurity = 0.429
15	12	on	→ Gini Impurity = 0.343
26.5	18	oui	→ Gini Impurity = 0.476
36.5	35	oui	→ Gini Impurity = 0.476
44	38	oui	→ Gini Impurity = 0.343
66.5	50	non	→ Gini Impurity = 0.429
	83	non	

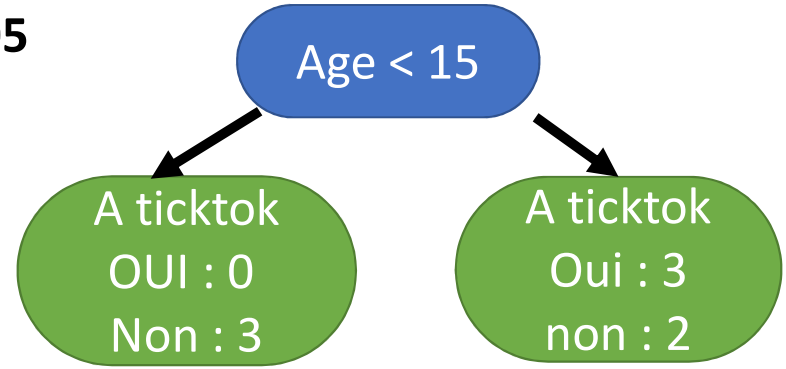
Gini Impurity



Gini Impurity = 0.405



Gini Impurity = 0.214



Gini Impurity = 0.343

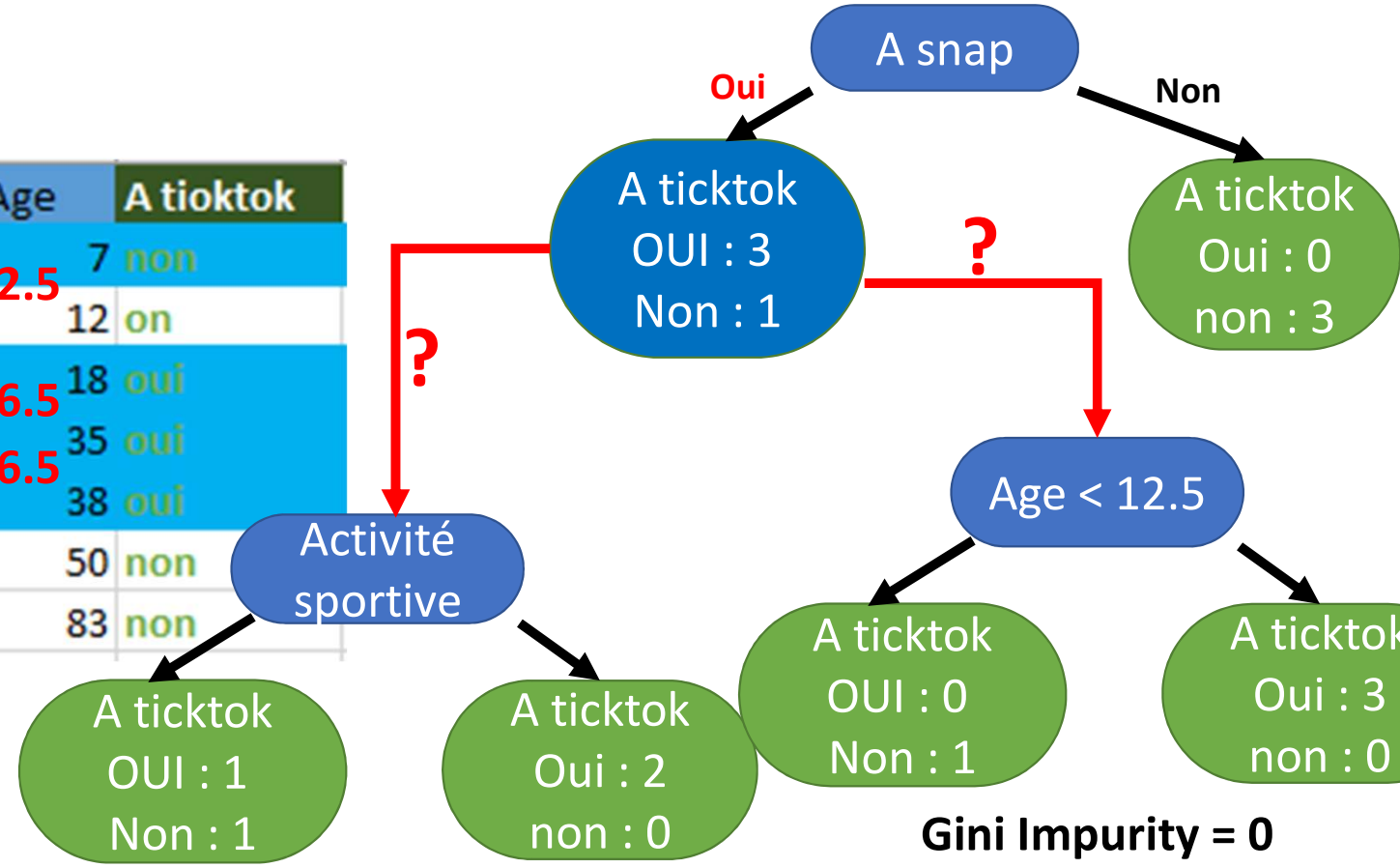
Arbres de décision

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non

12.5

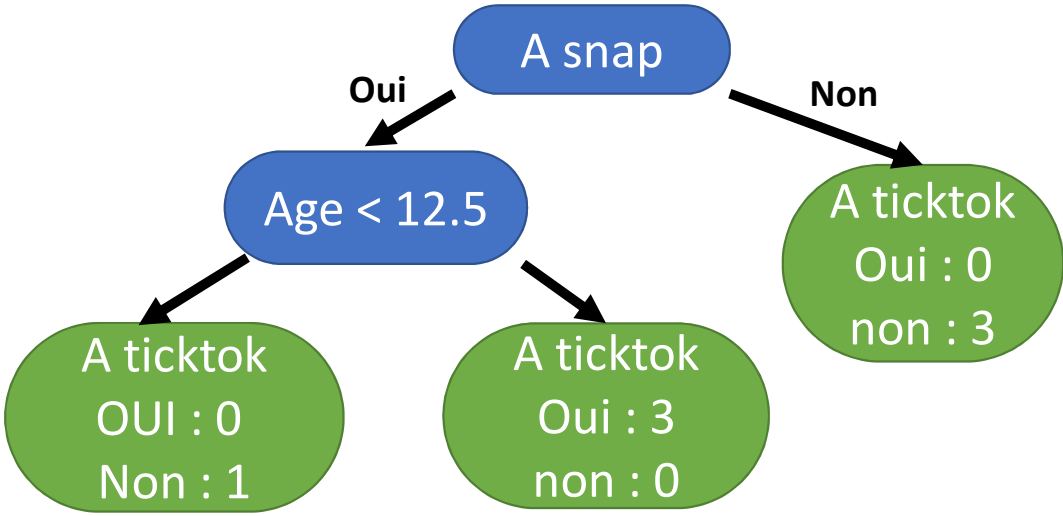
26.5

36.5



Arbres de décision

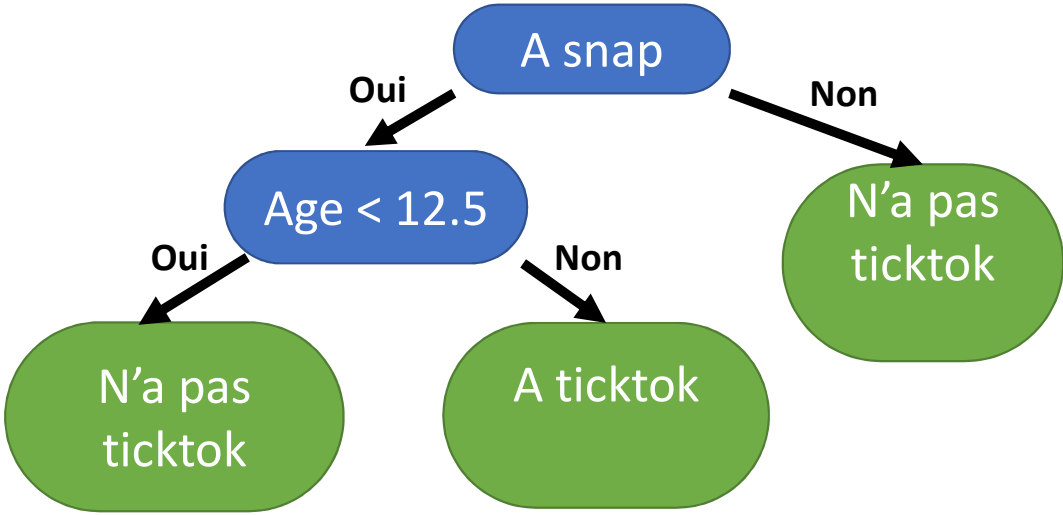
Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non



Arbres de décision

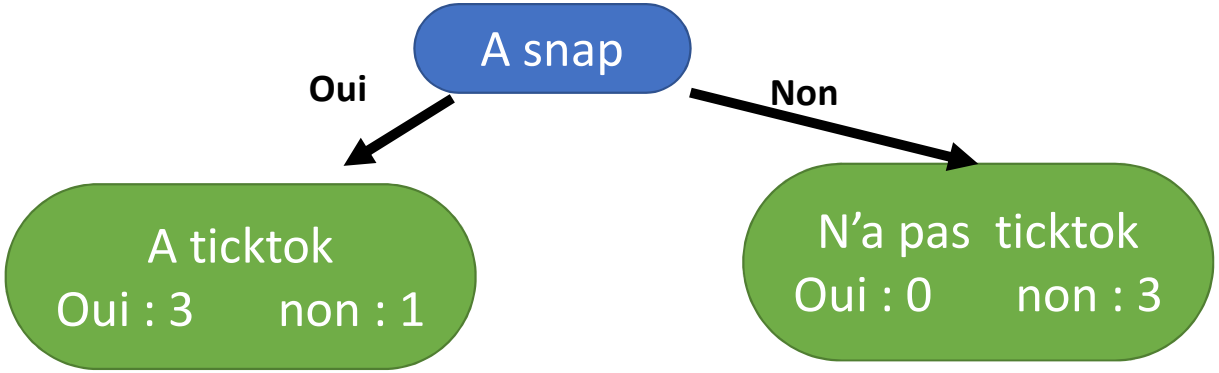
- Généralement la prédiction de chaque feuille est la valeur qui a le plus de votes.

Activité sportive	A snap	Age	A tioktok
oui	oui	7	non
oui	non	12	on
non	oui	18	oui
non	oui	35	oui
oui	oui	38	oui
oui	non	50	non
non	non	83	non



Overfitting

- Prunning.
- Put limites on how trees grow (set a minimum number of individuals in each leaf, this minimum can be chosen by crossvalidation).



Références

- WIREs Comput Stat 2013, 5:448–455. doi: 10.1002/wics.1278

- Statquest