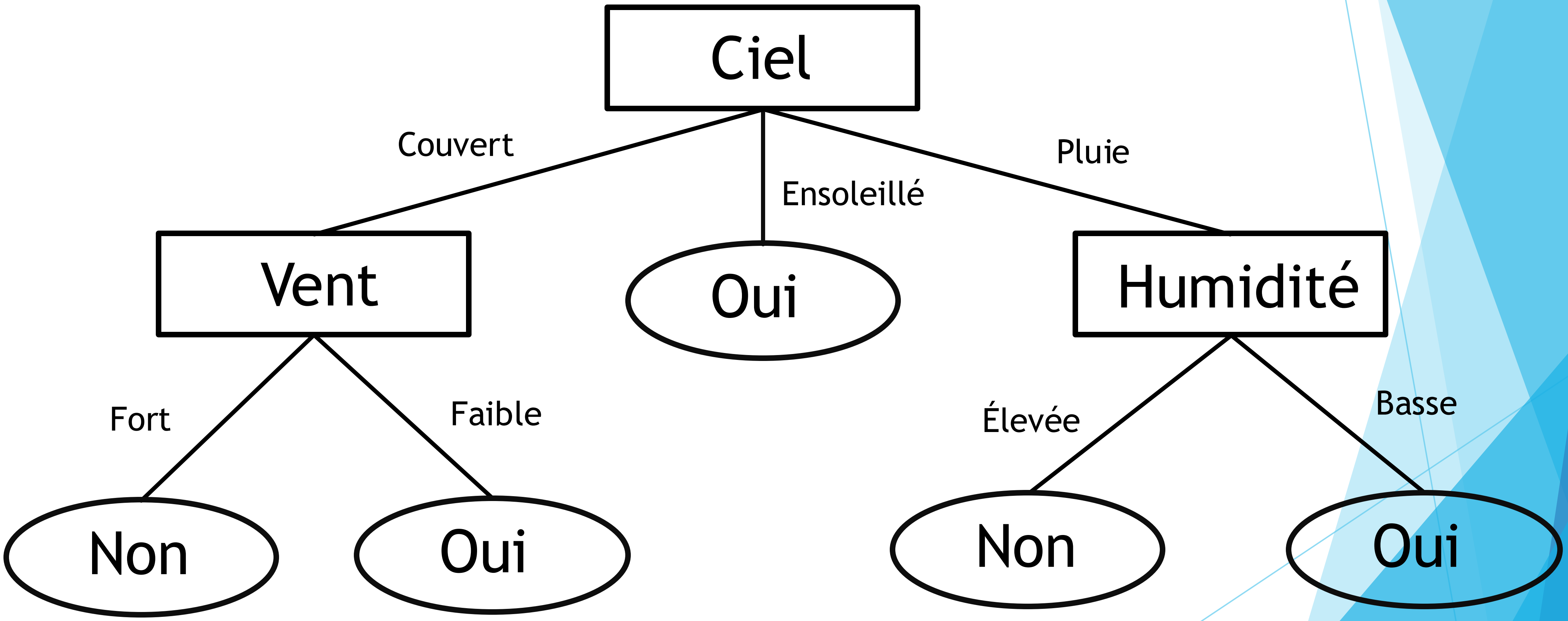
Arbre de décision

AI For You

Définition

- Les arbres de décision sont des algorithmes d'apprentissage automatique polyvalent qui peuvent effectuer des tâches de classification et de régression. Ce sont des algorithmes très puissants, capables de s'adapter à des ensembles de données complexes.

Vais-je jouer au tennis ?



Les points forts de l'arbre de décision

- L'une des nombreuses qualités des arbres de décision est qu'ils nécessitent très peu de préparation de données. En particulier, ils ne nécessitent pas du tout de mise à l'échelle ou de centrage des caractéristiques.

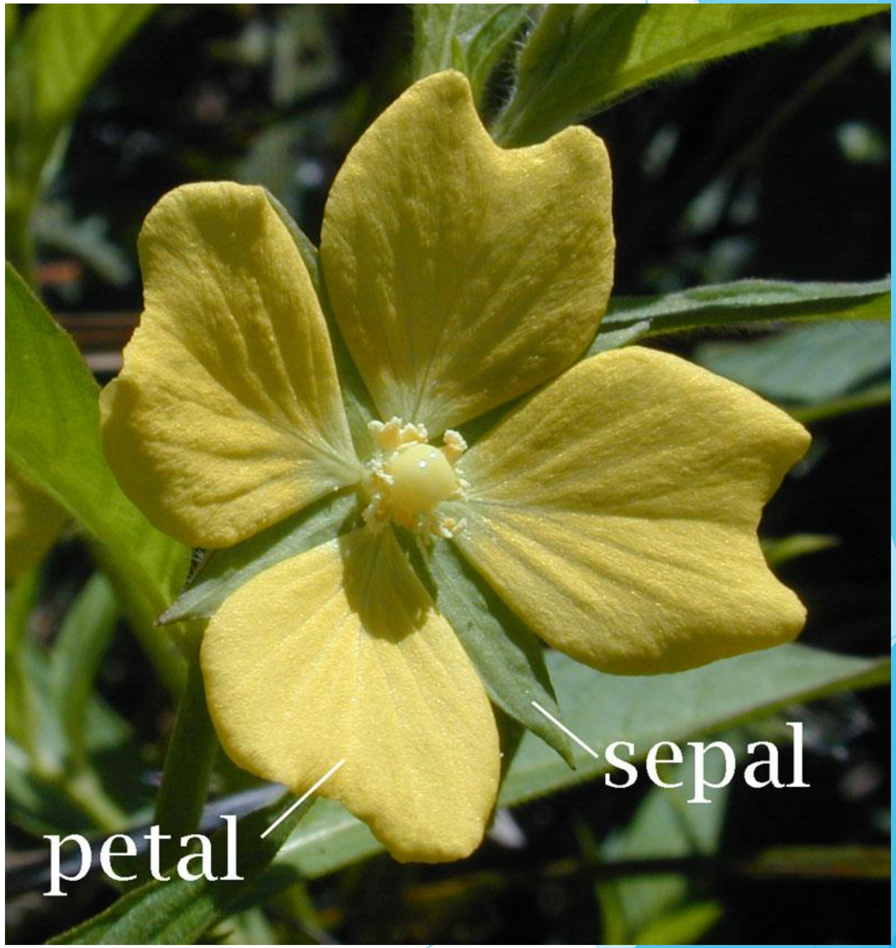
- Comme vous pouvez le voir, les arbres de décision sont assez intuitifs et leurs décisions sont faciles à interpréter, nous appelons ce genre de modèle : des boîtes blanches. En revanche, d'autres algorithmes comme le boosting ou les réseaux de neurones sont généralement considérés comme des modèles de boîte noire.

Arbre de classification

Le jeu de données

- Base de données iris

Sepal length	Sepal width	Petal length	Petal width	Espèce
5.1	3.5	1.4	0.2	I. setosa
7.0	3.2	4.7	1.4	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
⋮	⋮	⋮	⋮	⋮



Le jeu de données

- Base de données iris

Sepal length	Sepal width	Petal length	Petal width	Espèce
5.1	3.5	1.4	0.2	I. setosa
7.0	3.2	4.7	1.4	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
⋮	⋮	⋮	⋮	⋮

Setosa

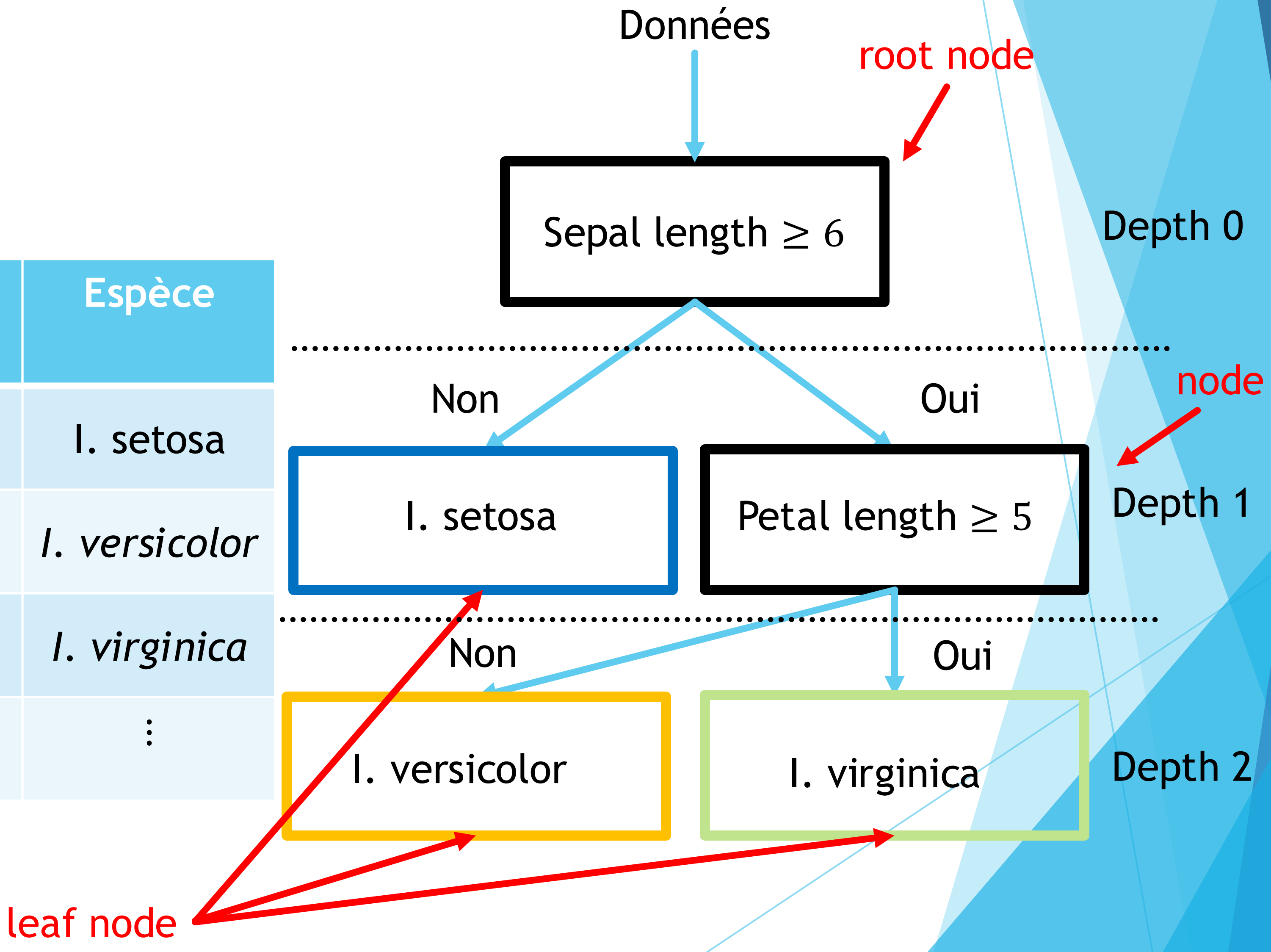
Versicolor

Virginica

Votre premier modèle d'arbre de décision

- Base de données iris

Sepal length	Sepal width	Petal length	Petal width	Espèce
5.1	3.5	1.4	0.2	I. setosa
7.0	3.2	4.7	1.4	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
⋮	⋮	⋮	⋮	⋮



Estimation de la probabilité d'appartenance

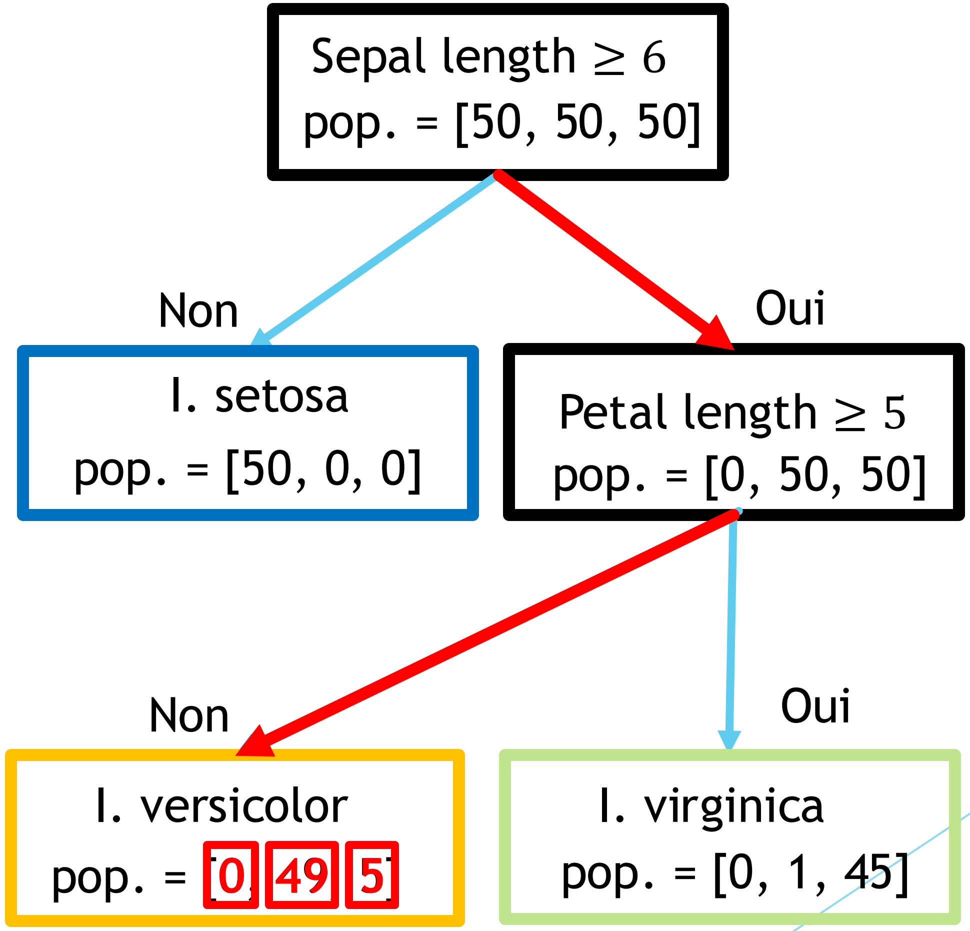
Un arbre de décision peut également estimer la probabilité qu'une observation appartienne à une classe k particulière.

Par exemple :

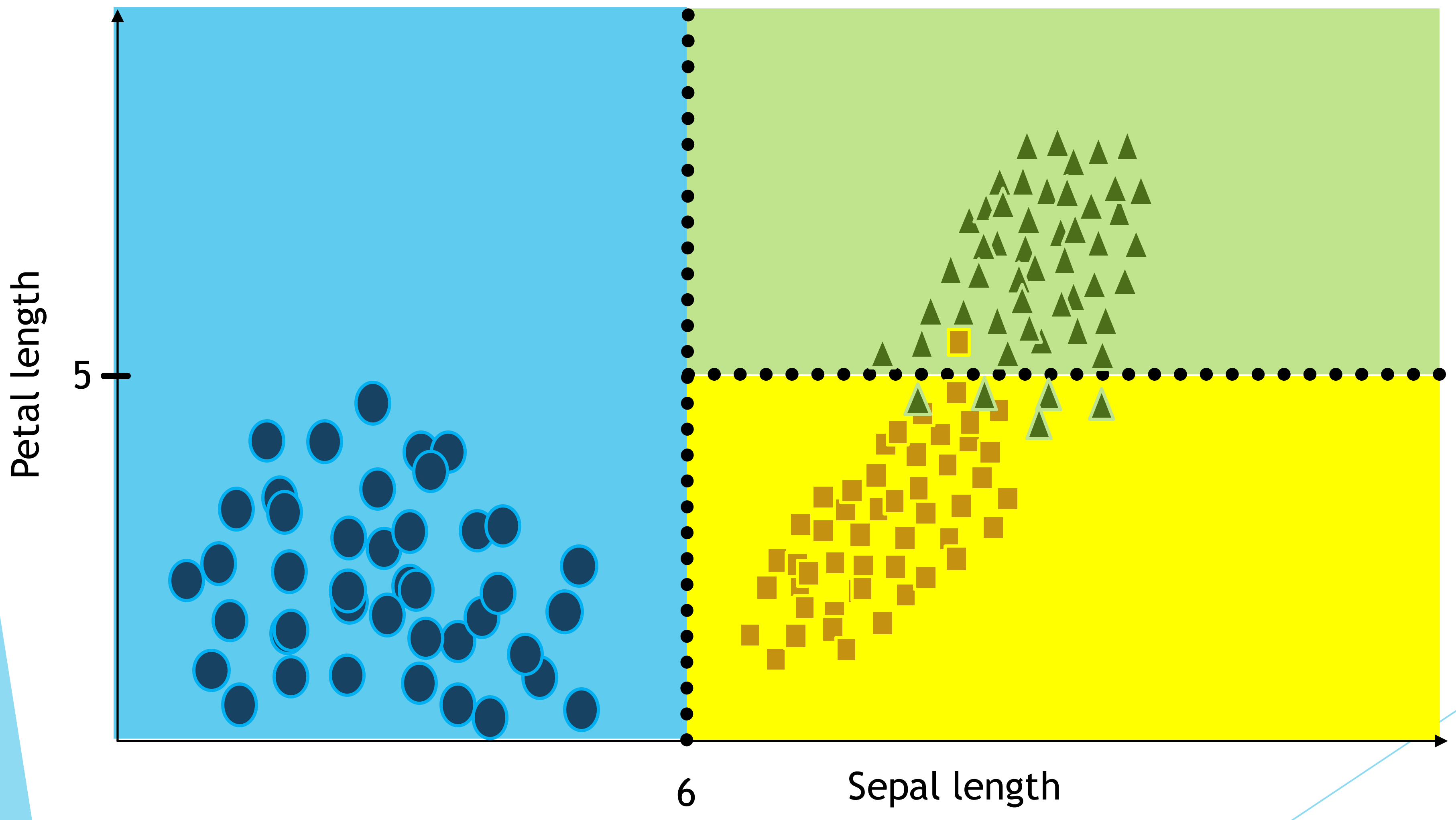
Nous voulons identifier une fleur avec un sépale de 6 cm et un pétale de 4 cm.

0/54 = 0% Setosa
49/54 = 90.7% Versicolor
5/54 = 9.3% Virginica

population = [nb setosa, nb versicolor, nb virginica]

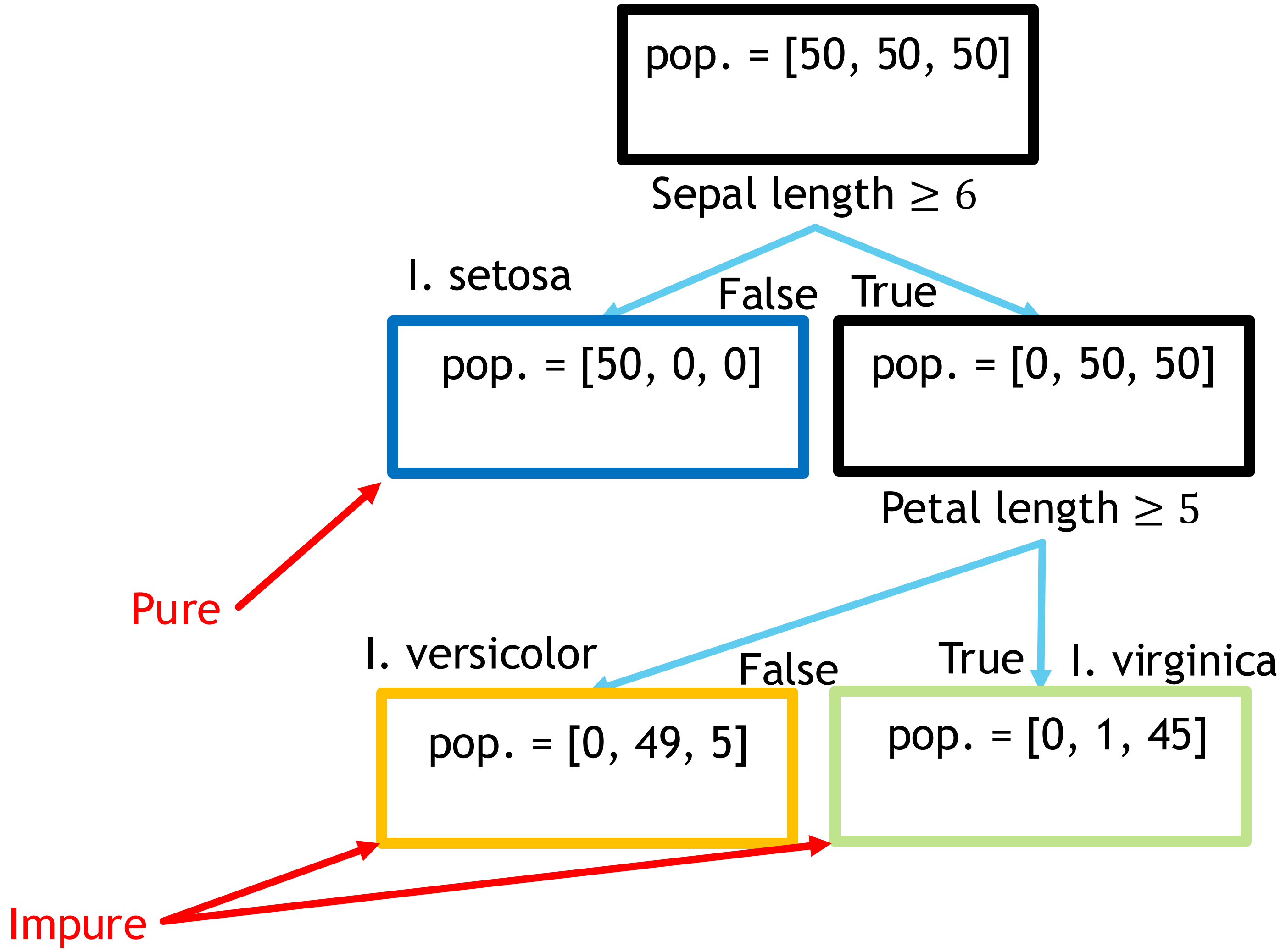


Frontière de décision



Pureté

population = [nb setosa, nb versicolor, nb virginica]



Indice Gini

population = [nb setosa, nb versicolor, nb virginica]

Comment créer la bonne inéquation qui conduira à une meilleure classification ?

$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

$p_{i,k}$ est le ratio du nombre d'individus de la classe k parmi la population du $i^{ème}$ noeud.

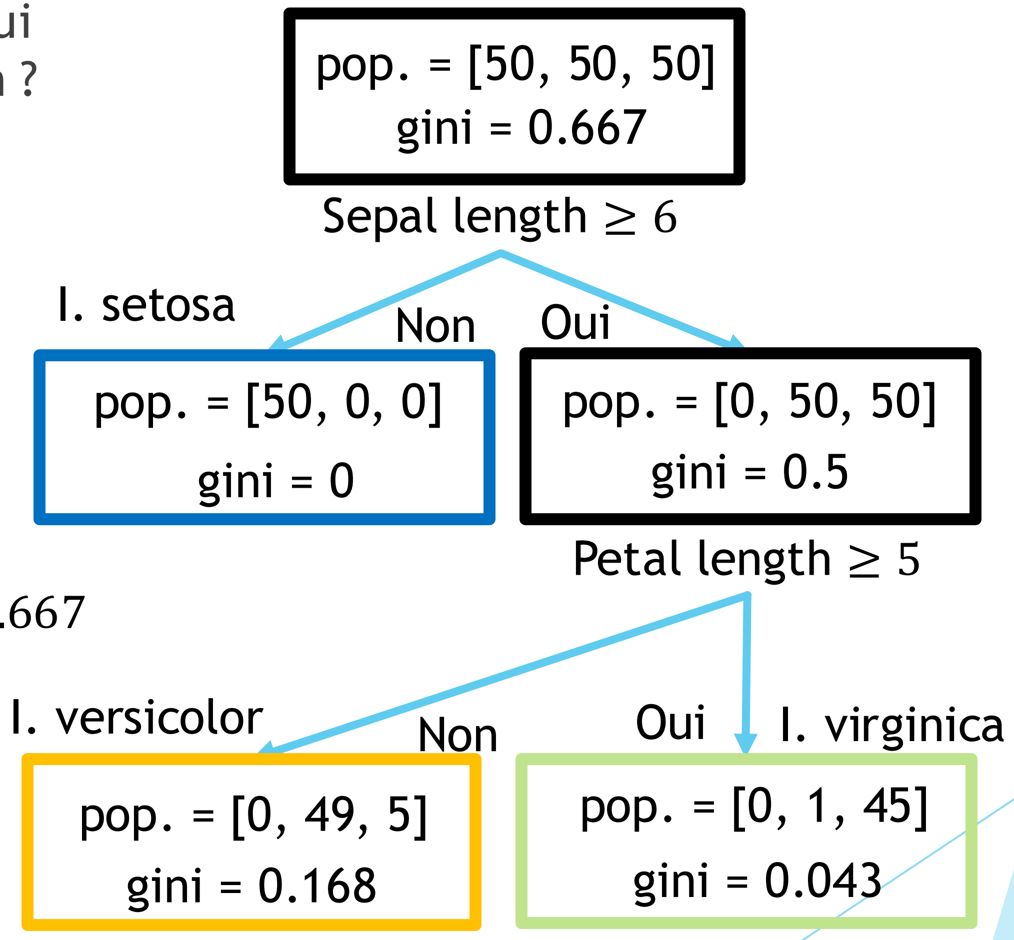
gini = $1 - \left(\frac{50}{150}\right)^2 - \left(\frac{50}{150}\right)^2 - \left(\frac{50}{150}\right)^2 = 0.667$

gini = $1 - \left(\frac{50}{50}\right)^2 - \frac{0}{50} - \frac{0}{50} = 0$

gini = $1 - \frac{0}{100} - \left(\frac{50}{100}\right)^2 - \left(\frac{50}{100}\right)^2 = 0.5$

gini = $1 - \frac{0}{54} - \left(\frac{49}{54}\right)^2 - \left(\frac{5}{54}\right)^2 = 0.168$

gini = $1 - \frac{0}{46} - \left(\frac{1}{46}\right)^2 - \left(\frac{45}{46}\right)^2 = 0.043$



Coût du nœud

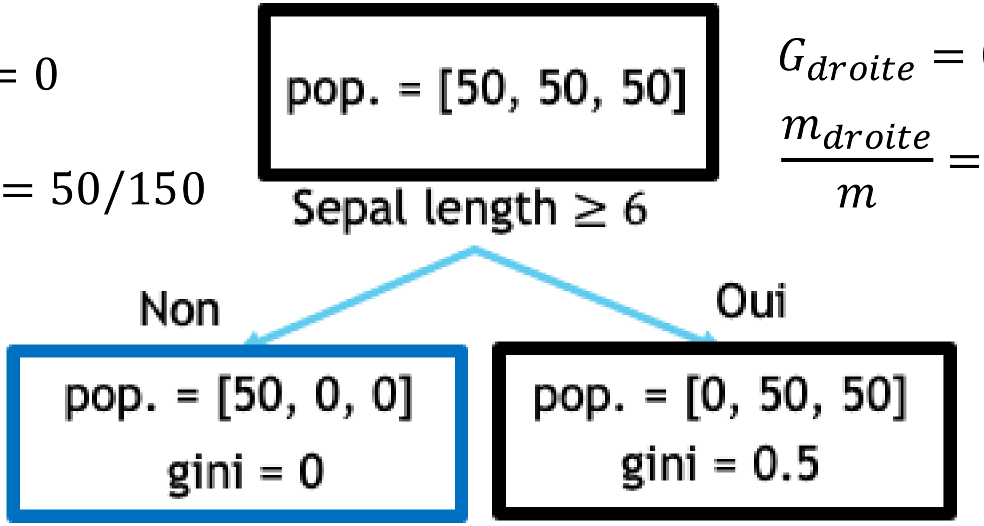
On veut calculer pour la pureté de notre nœud *k*

$$J(k) = \frac{m_{gauche}}{m} G_{gauche} + \frac{m_{droite}}{m} G_{droite}$$

$$Ou \begin{cases} G_{gauche/droite} \; mesure\ l'impureté\ du\ sous\ ensemble\ droite/gauche \\ m_{gauche/droite} \; est\ la\ proportion\ de\ notre\ population\ du\ sous\ ensemble\ droite/gauche \end{cases}$$

$G_{gauche} = 0$

$\frac{m_{gauche}}{m} = 50/150$



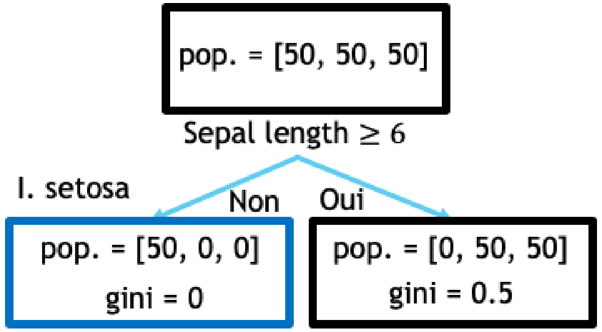
$G_{droite} = 0.5$

$\frac{m_{droite}}{m} = 100/150$

$$J(0) = \frac{50}{150} 0 + \frac{100}{150} 0.5 = 0.33$$

Choisir les noeuds

1)



$$G_{gauche} = 0$$

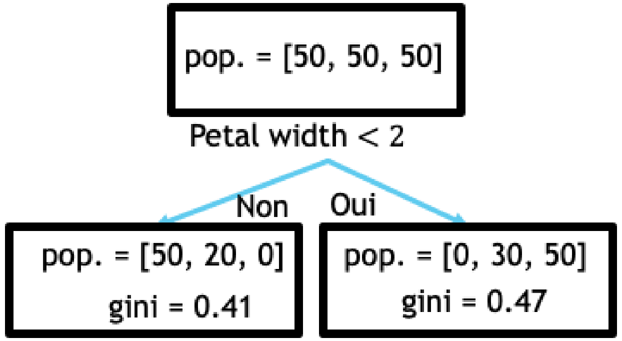
$$m_{gauche} = 50/150$$

$$G_{droite} = 0.5$$

$$m_{droite} = 100/150$$

$$J(0) = \frac{50}{150}\,0 + \frac{100}{150}\,0.50 = \boxed{\underline{0.33}}$$

2)



$$G_{gauche} = 0.41$$

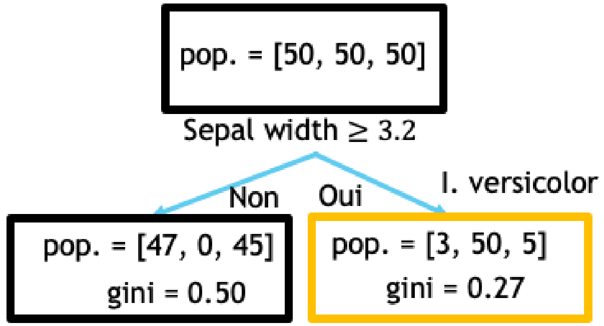
$$m_{gauche} = 70/150$$

$$G_{droite} = 0.47$$

$$m_{droite} = 80/150$$

$$J(0) = \frac{70}{150}\,0.41 + \frac{80}{150}\,0.47 = \underline{0.44}$$

3)



$$G_{gauche} = 0.60$$

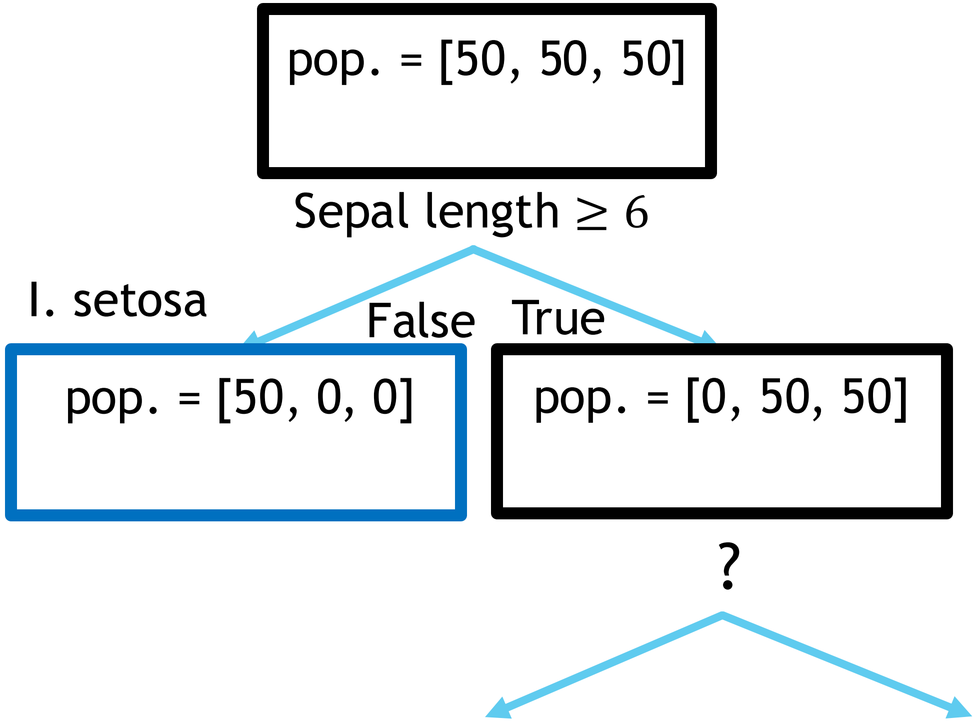
$$m_{gauche} = 92/150$$

$$G_{droite} = 0.27$$

$$m_{droite} = 58/150$$

$$J(0) = \frac{92}{150}\,0.0 + \frac{58}{150}\,0.27 = \underline{0.41}$$

Construction d'un arbre



Arbre de régression

Le jeu de données

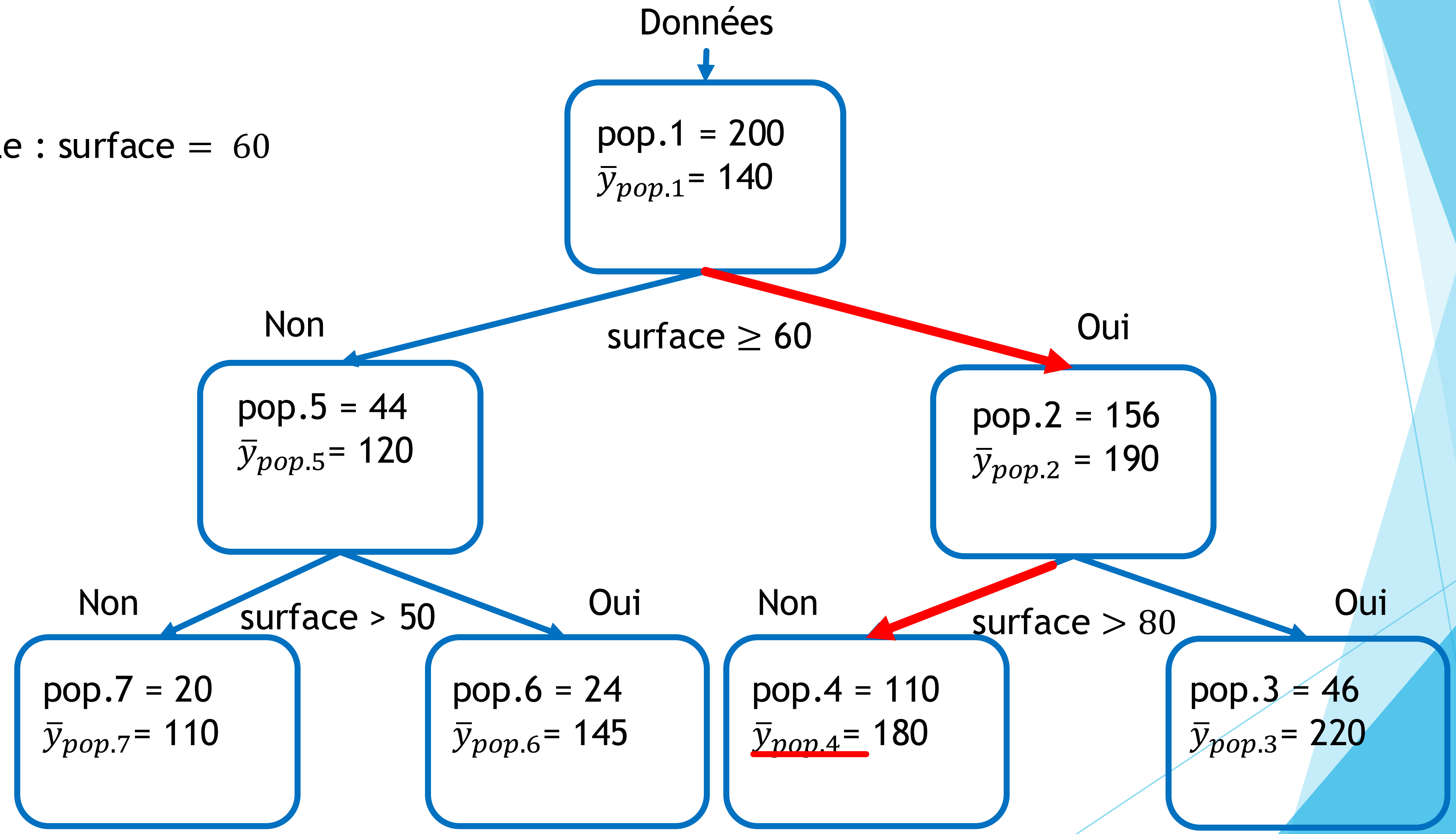
- Base de données prix de maison

Nb pièces	surface	Garage	Année	Prix (k€)
3	72	1	2017	180
2	58	1	2010	*140*
3	76	0	1998	*160*
⋮	⋮	⋮	⋮	⋮

Arbre de régression

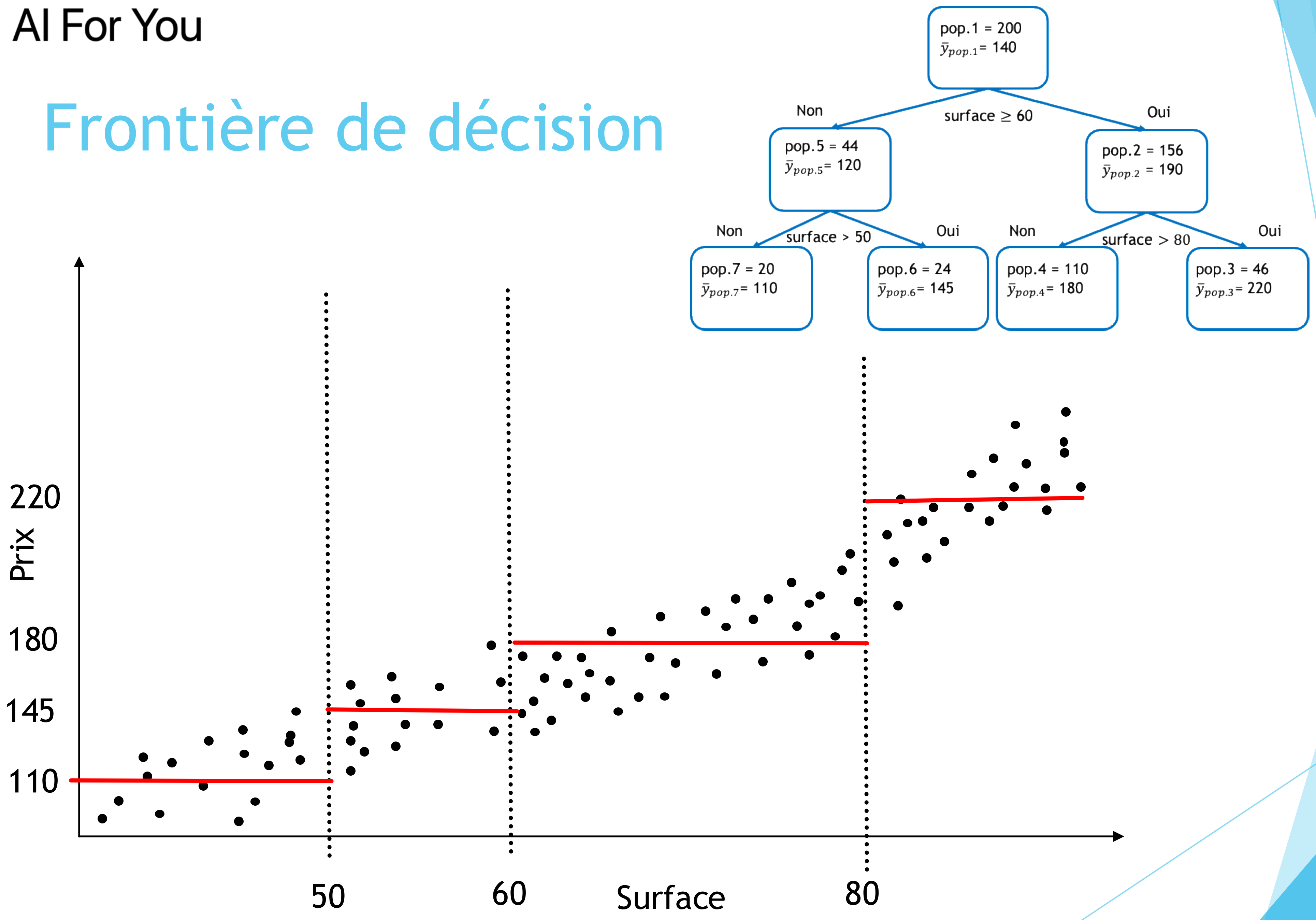
$moyenne = \bar{y}$

Exemple : surface $= 60$



AI For You

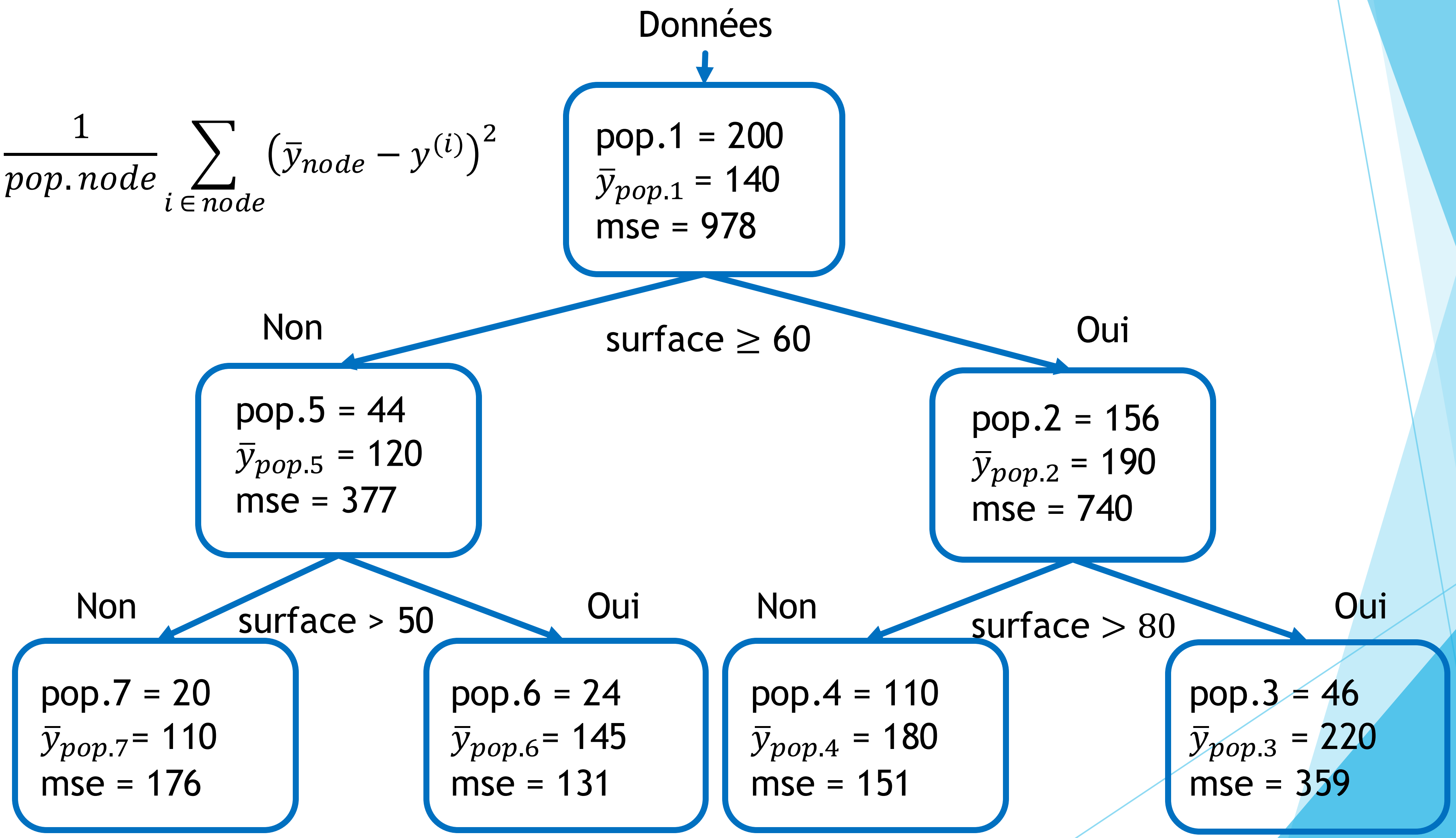
Frontière de décision



Pureté

$$moyenne = \bar{x}$$

$$MSE = \frac{1}{pop.node} \sum_{i \in node} \left(\bar{y}_{node} - y^{(i)}\right)^2$$



Coût du nœud

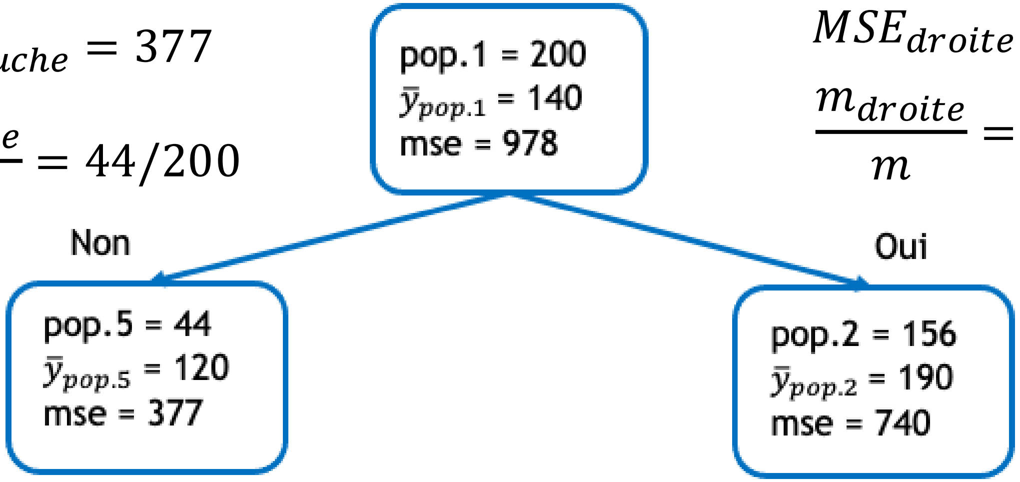
On veut calculer pour la pureté de notre nœud *k*

$$J(k) = \frac{m_{gauche}}{m} MSE_{gauche} + \frac{m_{droite}}{m} MSE_{droite}$$

$$Où \begin{cases} MSE_{gauche/droite} = \dfrac{1}{pop.node} \displaystyle\sum_{i \in node} \left(\hat{y}_{node} - y^{(i)}\right)^2 \\ \hat{y}_{node} = \dfrac{1}{m_{node}} \displaystyle\sum_{i \in node} y^{(i)} \end{cases}$$

$$MSE_{gauche} = 377$$

$$\frac{m_{gauche}}{m} = 44/200$$



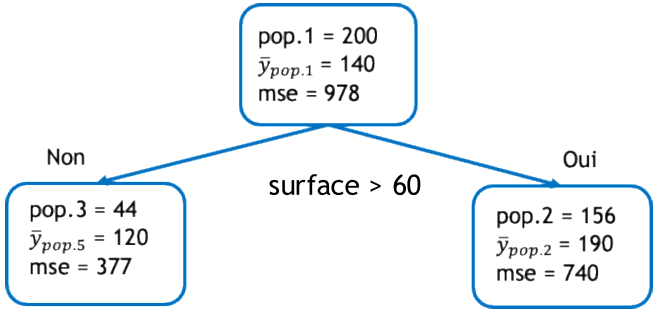
$$MSE_{droite} = 740$$

$$\frac{m_{droite}}{m} = 156/200$$

$$J(0) = \frac{44}{200} 377 + \frac{156}{200} 740 = 660$$

Choisir les noeuds

1)



$$MSE_{gauche} = 377$$

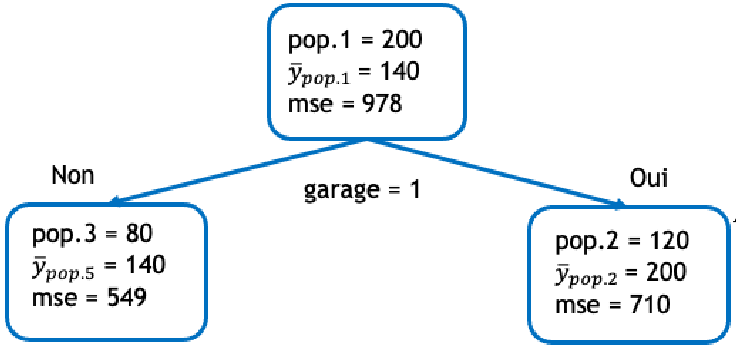
$$\frac{m_{gauche}}{m} = 44/200$$

$$MSE_{droite} = 740$$

$$\frac{m_{droite}}{m} = 156/200$$

$$J(0) = \frac{44}{200}\,377 + \frac{156}{200}\,740 = \underline{660}$$

2)



$$MSE_{gauche} = 549$$

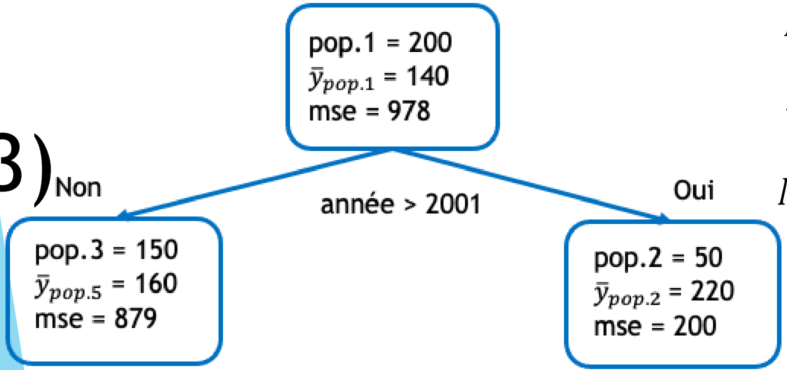
$$\frac{m_{gauche}}{m} = 80/200$$

$$MSE_{droite} = 710$$

$$\frac{m_{droite}}{m} = 120/200$$

$$J(0) = \frac{80}{200}\,549 + \frac{120}{200}\,710 = \boxed{\underline{646}}$$

3)



$$MSE_{gauche} = 879$$

$$\frac{m_{gauche}}{m} = 150/200$$

$$MSE_{droite} = 200$$

$$\frac{m_{droite}}{m} = 50/200$$

$$J(0) = \frac{150}{200}\,879 + \frac{50}{200}\,200 = \underline{709}$$

Des solutions raisonnablement bonne

- L'algorithme CART est un algorithme gourmand : il recherche avidement une répartition optimale au niveau supérieur, puis répète le processus à chaque niveau. Il ne vérifie pas si la scission conduira ou non à une impureté aussi faible que possible plusieurs niveaux plus bas. Un algorithme gourmand produit souvent une solution raisonnablement bonne, mais il n'est pas garanti que ce soit la solution optimale. Parce que l'arbre optimal est connu pour être un problème NP-Complet.

Limitation de l'arbre de décision

- Les arbres de décision aiment les limites de décision orthogonales, ce qui les rend sensibles à la rotation des ensembles d'entraînement.

- Ils sont également instables car ils sont très sensibles à de petites variations des données d'entraînement

Le random forest

- Le random forest vient combler certaines limites des arbres de décision.