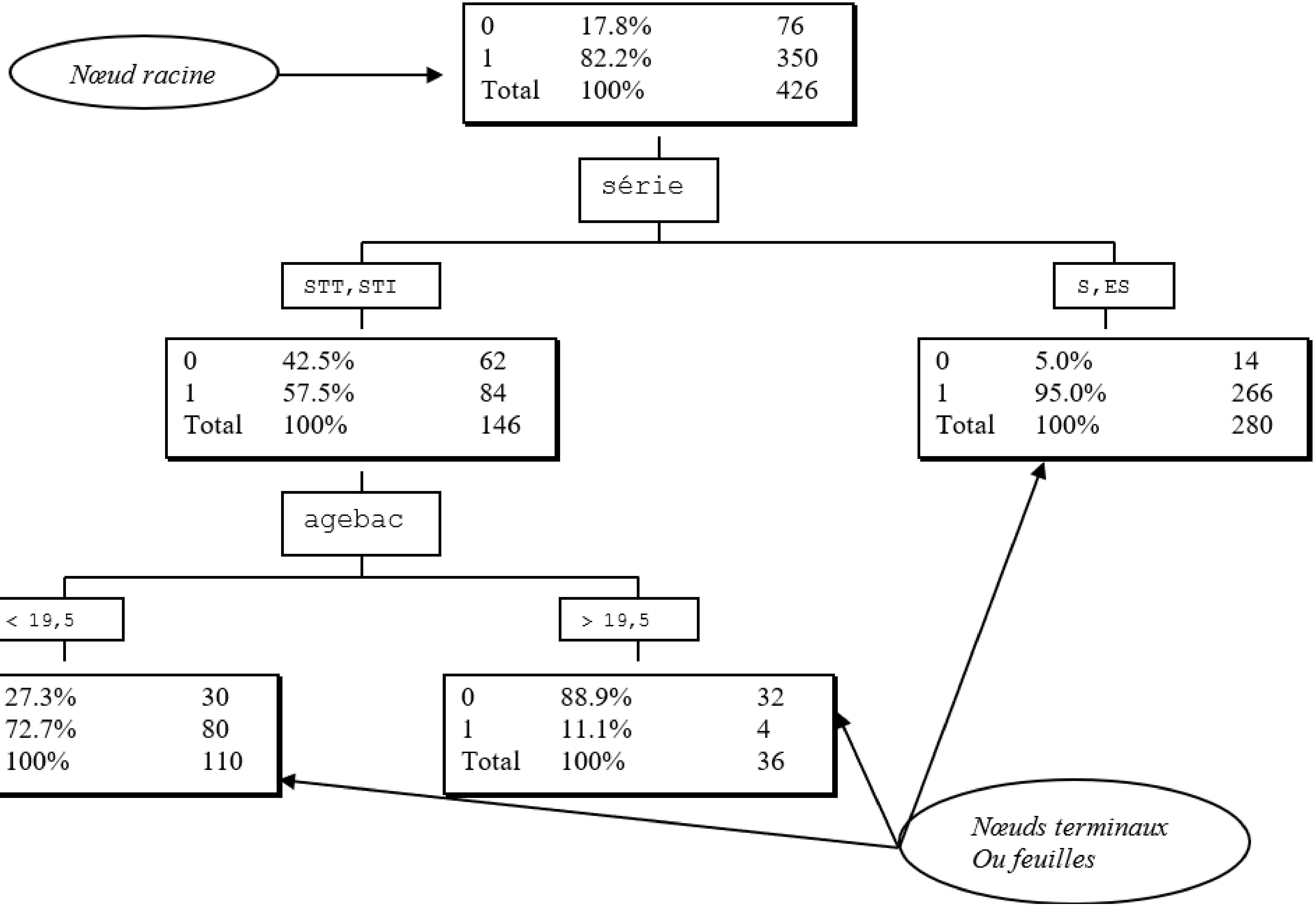
Les arbres de décision

Exemple : Réussite en deux ans en IUT

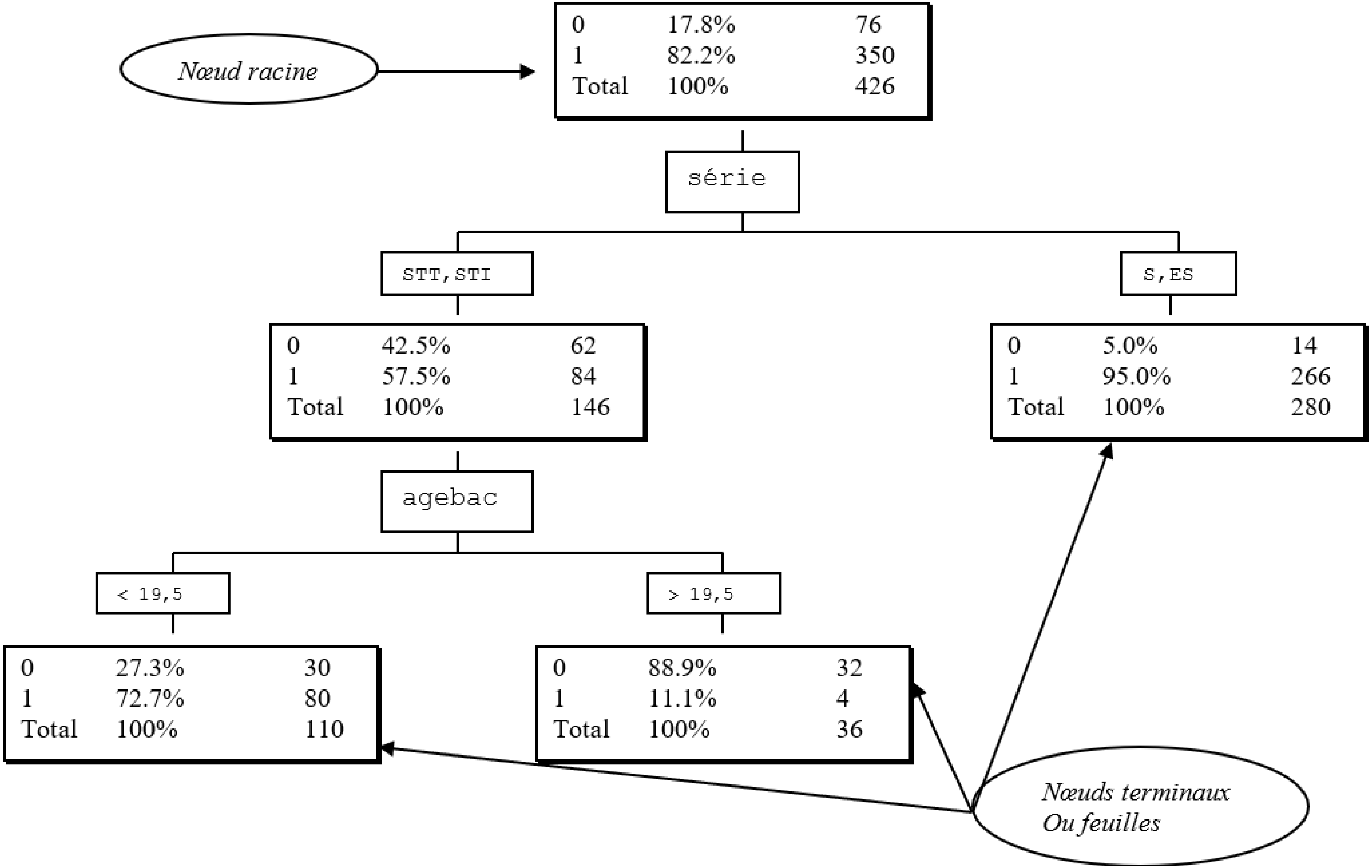


Arbres de décision

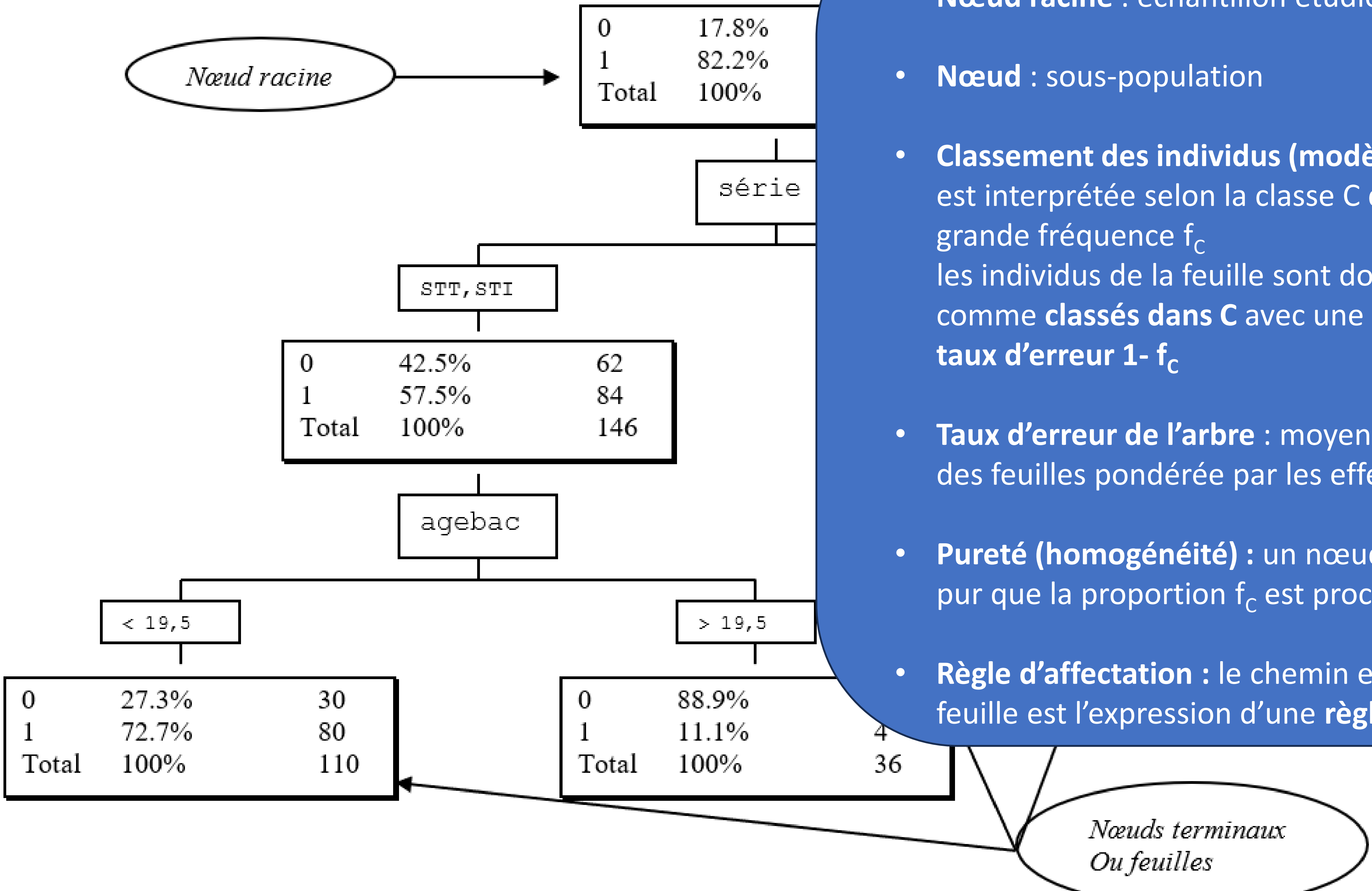
- Technique d'apprentissage supervisé
- $\boldsymbol{Y}$ en fonction de variables explicatives $\boldsymbol{X_1}, \boldsymbol{X_2} \ldots \boldsymbol{X_p}$.
  - Si Y est quantitative continue : arbre de régression
  - $\boldsymbol{X_1}, \boldsymbol{X_2} \ldots \boldsymbol{X_p}$ Variables qualitatives ou quantitatives

La technique des arbres de décision aboutit à une représentation graphique qui met en évidence un ensemble de règles aisément interprétables.

Exemple : Réussite en deux ans en IUT



Exemple : Réussite en deux ans en IUT



VOCABULAIRE

- **Nœud racine** : échantillon étudié
- **Nœud** : sous-population
- **Classement des individus (modèle) :** chaque feuille est interprétée selon la classe C de Y qui la plus grande fréquence f_C
  les individus de la feuille sont donc considérés comme **classés dans C** avec une **probabilité f_C** et **un taux d'erreur 1- f_C**
- **Taux d'erreur de l'arbre** : moyenne des taux d'erreur des feuilles pondérée par les effectifs des feuilles
- **Pureté (homogénéité) :** un nœud est d'autant plus pur que la proportion f_C est proche de 1
- **Règle d'affectation :** le chemin entre la racine et une feuille est l'expression d'une **règle**.

Exemple : Réussite en deux ans en IUT

Nœud racine

0	17.8%	
1	82.2%	
Total	100%	

série

Vocabulaire

- **Nœud racine** : échantillon étudié
- **Nœud** : sous-population
- **Classement des individus :** chaque feuille est … C de Y qui la plus grande
- … nœud est d’autant plus … proche de 1 ou de 0.
- … min entre la racine et une … **règle**.

Avantages

- **Règles lisibles**, détection de « niches »
- **Peu d’hypothèses** sur les données.
- Robuste vis-à-vis de données **erronées, aberrantes** ou **manquantes**

Inconvénient

- **Sensibilité** à de légères modifications
- Arbre **optimal** ?

< 19,5

0	27.3%	30
1	72.7%	80
Total	100%	110

> 19,5

0	88.9%	32
1	11.1%	4
Total	100%	36

Nœuds terminaux
Ou feuilles

Comment obtient-on un arbre de décision

Chaque étape correspond à la division d'un nœud

- **Pour diviser un nœud :**
  On recherche la variable qui **explique** le mieux la variable cible : la plus **discriminante** si la variable cible est nominale (la **plus liée** au phénomène décrit par Y, si Y quantitative)

- Cette variable définit une division de la population en **deux sous-populations** (nœuds).

- On réitère le processus sur les sous-populations obtenues en recherchant une seconde variable et ainsi de suite...

Construction d’un arbre binaire

Cas particulier d’une variable binaire

X1	X2	X3	Y
27,4	Bayonne	Très Bien	1
30,8	Anglet	Assez Bien	1
26,7	Anglet	Bien	0
24,7	Anglet	Bien	0
24,7	Anglet	Bien	0
26,3	Biarritz	Assez Bien	0
29,2	Anglet	Très Bien	1
30	Biarritz	Assez Bien	1
23,6	Bayonne	Bien	1
24,5	Biarritz	Bien	0
25,2	Anglet	Très Bien	0
24,5	Bayonne	Bien	0
27,7	Bayonne	Très Bien	1
26,8	Anglet	Très Bien	0
26,1	Biarritz	Assez Bien	0
30,4	Bayonne	Bien	1
27,8	Bayonne	Très Bien	0
26,3	Bayonne	Bien	0
24,1	Anglet	Très Bien	0
28,9	Bayonne	Bien	0
25,8	Bayonne	Assez Bien	1
24,5	Anglet	Bien	0
28,1	Bayonne	Assez Bien	0
29,5	Anglet	Très Bien	1
28,9	Bayonne	Bien	0
26,7	Bayonne	Bien	1
26,1	Bayonne	Bien	0

Combien de divisions possibles selon le type de la variable ?

50 individus (lignes) et 4 variables :

- **X1 :** variable quantitative continue prenant **34 valeurs distinctes**
- **X2 :** variable qualitative nominale prenant **3 modalités :** 'Bayonne', 'Anglet' et 'Biarritz'
- **X3 :** variable qualitative ordinale prenant **3 modalités** : 'Assez Bien', 'Bien' et 'Très Bien'
- **Y :** variable cible binaire

Combien a-t-on de divisions possibles pour chacune des variables **X1,X2** et **X3** ?

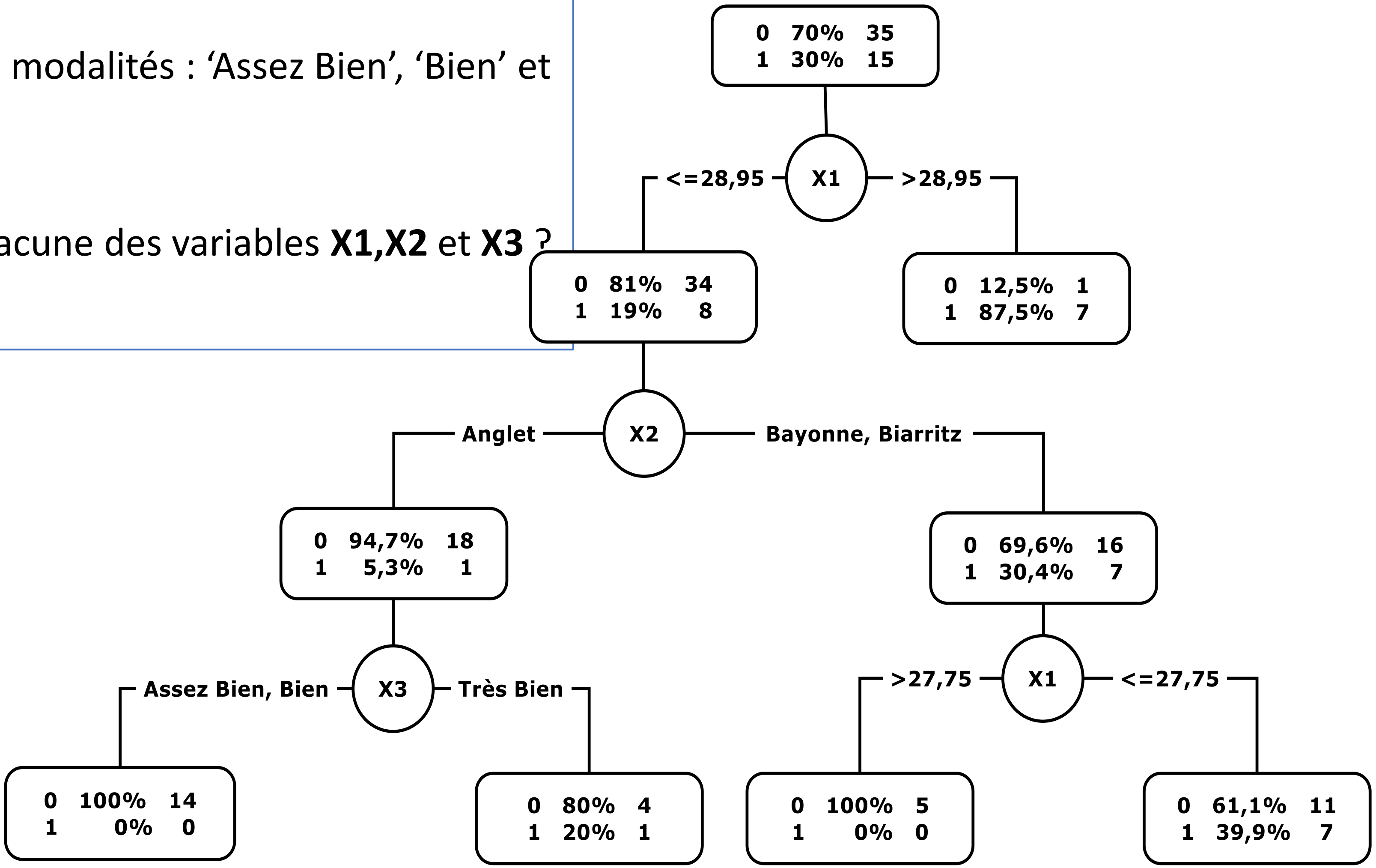
Combien de divisions possibles selon le type de la variable ?

X1	X2	X3	Y
27,4	Bayonne	Très Bien	1
30,8	Anglet	Assez Bien	1
30,4	Bayonne	Bien	1
27,8	Bayonne	Très Bien	0
26,3	Bayonne	Bien	0
24,1	Anglet	Très Bien	0
28,9	Bayonne	Bien	0
25,8	Bayonne	Assez Bien	1
24,5	Anglet	Bien	0
28,1	Bayonne	Assez Bien	0
29,5	Anglet	Très Bien	1
28,9	Bayonne	Bien	0
26,7	Bayonne	Bien	1
26,1	Bayonne	Bien	0

50 individus (lignes) et 4 variables :

- **X1 :** variable quantitative continue prenant 34 valeurs distinctes
- **X2 :** variable qualitative nominale prenant 3 modalités : ‘Bayonne’, ‘Anglet’ et ‘Biarritz’
- **X3 :** variable qualitative ordinale prenant 3 modalités : ‘Assez Bien’, ‘Bien’ et ‘Très Bien’
- **Y :** variable cible binaire

Combien a-t-on de divisions possibles pour chacune des variables **X1,X2** et **X3** ?



Nombre de divisions possibles en deux classes pour une variable X_i

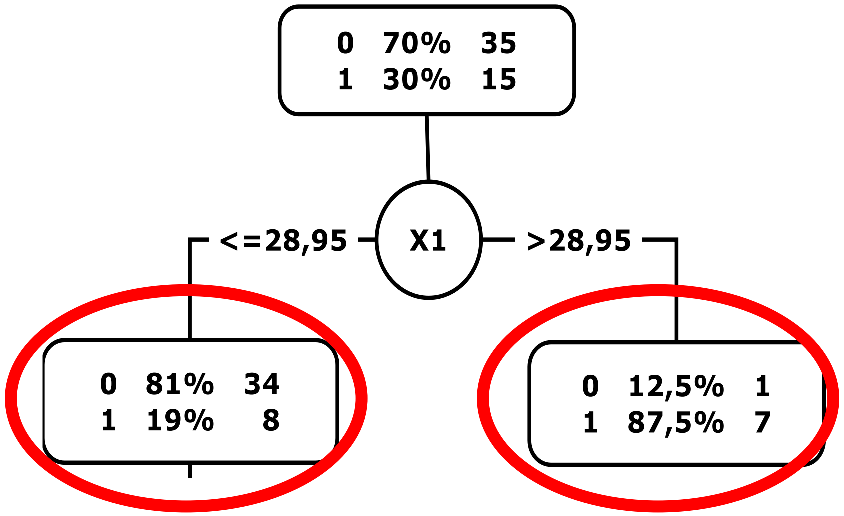
$\rightarrow$ **Si $\boldsymbol{X_i}$ est qualitative à $\boldsymbol{q}$ modalités** :

- $q-1$ divisions possibles si X_i est ordinale
- $2^{q-1}-1$ divisions possibles si X_i est nominale

$\rightarrow$ **Si $\boldsymbol{X_i}$ est quantitative et prend $\boldsymbol{n}$ valeurs** : $n-1$ divisions possibles

Division la plus « intéressante » ?

Celle qui donnera les nœuds les plus « différents » (discrimination) et donc, pour une variable binaire, **les plus homogènes** par rapport à la cible



Pour **mesurer l’homogénéité**, plusieurs critères : Khi-deux, entropie ou l’indice de Gini.

Indice de Gini (indice de Gini-Simpson, indice de diversité de Gini)

- Indicateur de « **pureté** » d’un nœud,
- **Probabilité** que deux individus choisis au hasard (avec remise) dans le nœud n’appartiennent pas à la même catégorie.

Indice de Gini (indice de Gini-Simpson*, indice de diversité* de Gini)

- Indicateur de « **pureté** » d'un nœud,
- **Probabilité** que deux individus choisis au hasard (avec remise) dans le nœud n'appartiennent pas à la même catégorie.

*On utilise les termes de **Gini-Simpson** ou indice **de diversité** pour distinguer cet indice d'un indice de Gini beaucoup plus connu, qui mesure les inégalités de répartition d'une variable (richesse, revenu) dans une population

Indice de Gini (indice de Gini-Simpson, indice de diversité de Gini)

- Indicateur de « **pureté** » d'un nœud,
- **Probabilité** que deux individus choisis au hasard (avec remise) dans le nœud n'appartiennent pas à la même catégorie.

Calcul pour deux modalités ?

On note $\mathbf{f_0}$ la fréquence de 0 dans le nœud et $\mathbf{f_1}$ la fréquence de 1 dans le nœud $(f_0 + f_1 = 1)$

On note les événements suivants :

- I : « deux individus choisis au hasard (avec remise) dans le nœud n'appartiennent pas à la même catégorie »
- C_i^j : « le $i^{\text{ème}}$ individu appartient à la catégorie j»

Indice de Gini

$$\boldsymbol{Gini = f_0 f_1 + f_1 f_0 = 1 - f_1^2 + f_0^2}$$

Généralisation à n modalités

$$\boldsymbol{Gini = \sum_{i \neq j} f_i f_j = 1 - \sum_{i=1}^{p} f_i^2}$$

Indice de Gini

$$\boldsymbol{Gini = f_0 f_1 + f_1 f_0 = 1 - f_1^2 + f_0^2}$$

Petit exercice d'analyse : dans le cas de deux modalités, notons x la fréquence f_0.

a. Exprimer l'indice de Gini sous forme d'une fonction de x, $G(x)$
b. Étudier les variations de G sur $[0,1]$. Quel est le minimum et le maximum de G sur $[0,1]$? Pour quelles valeurs de x sont-ils atteints ?

Indice de Gini

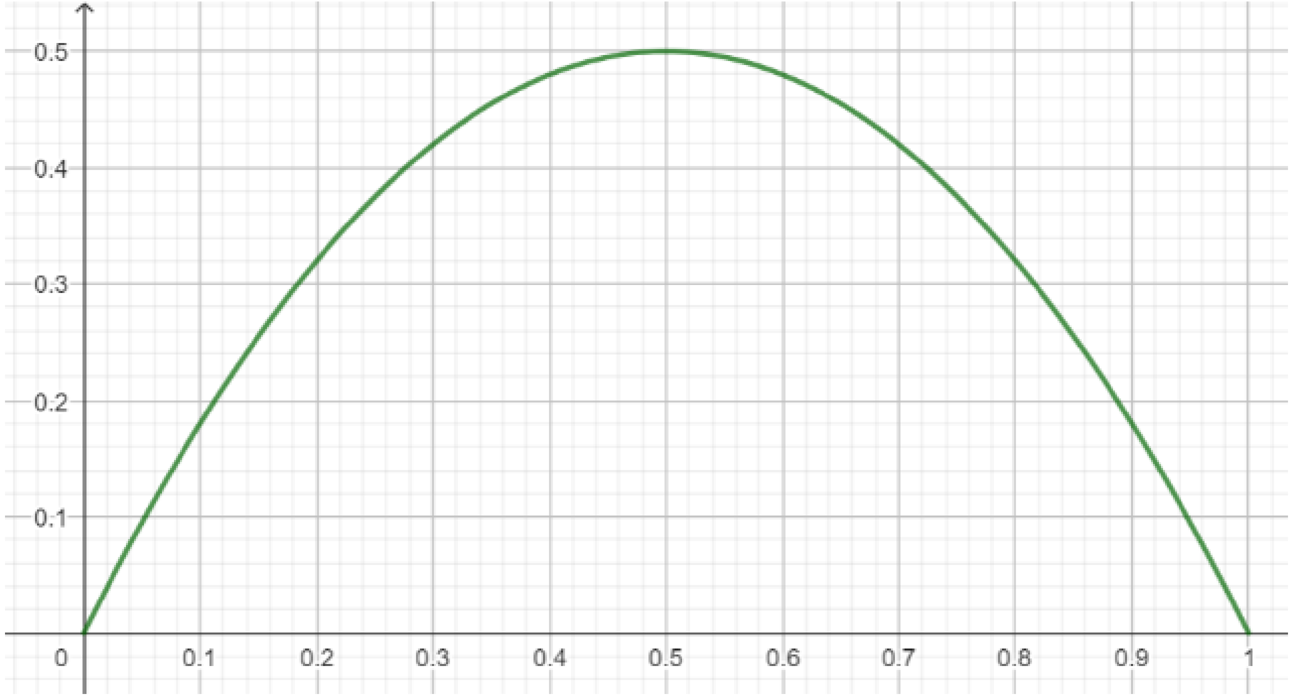
$$Gini = f_0 f_1 + f_1 f_0 = 1 - f_1^2 + f_0^2$$

Petit exercice d'analyse : dans le cas de deux modalités, notons x la fréquence f_0.

a. Exprimer l'indice de Gini sous forme d'une fonction de x, $G(x)$

b. Étudier les variations de G sur $[0,1]$. Quel est le minimum et le maximum de G sur $[0,1]$? Pour quelles valeurs de x sont-ils atteints ?

$$G(x) = 2x(1-x) = 2x - 2x^2$$



Étapes de construction de l’arbre

A partir du nœud racine,

- **Pour chaque variable X_i :**
    - On passe en revue **toutes les divisions** possibles en 2 classes de X_i.
    - On retient la division qui définit les 2 nœuds **les plus homogènes**, en utilisant par exemple l’indice de Gini pour mesurer l’homogénéité.

      Plus précisément, on calcule une fonction Gain (de pureté) définie par :

Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

On a choisi la variable (et sa division) optimisant l’homogénéité des nœuds, on divise et on obtient deux nouveaux nœuds.

- **On réitère le processus sur les nœuds obtenus**

Arrêt du processus

Exemples de critères d’arrêt :

- L’effectif de chaque nœud est inférieur à un seuil fixé ;
- Ou la profondeur de l’arbre a atteint une limite fixée ;
- Ou le nombre de feuilles a atteint un maximum fixé ;
- Ou la qualité de l’arbre n’augmente plus de façon suffisante ;
- Ou plus aucune division n’est possible…

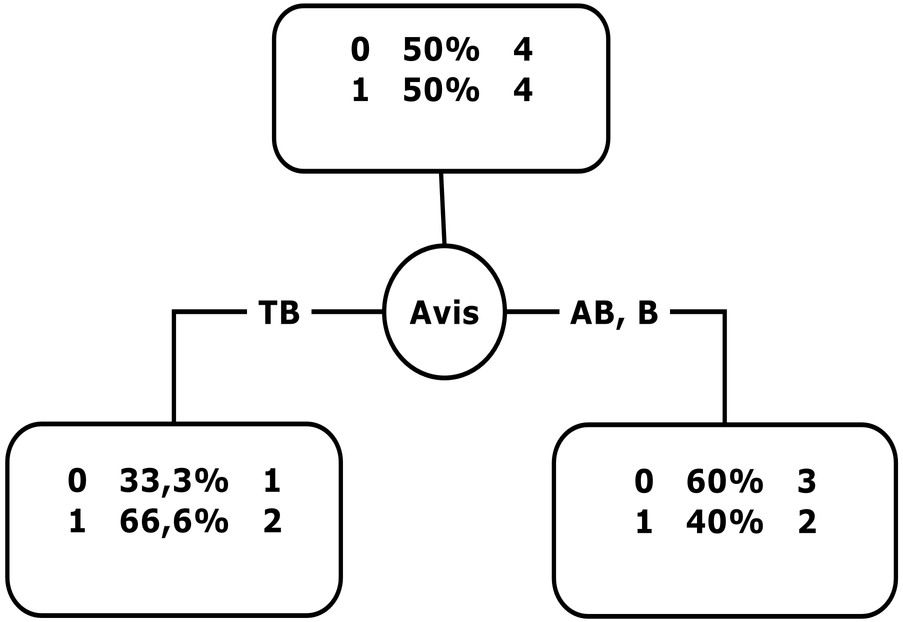
Exercice

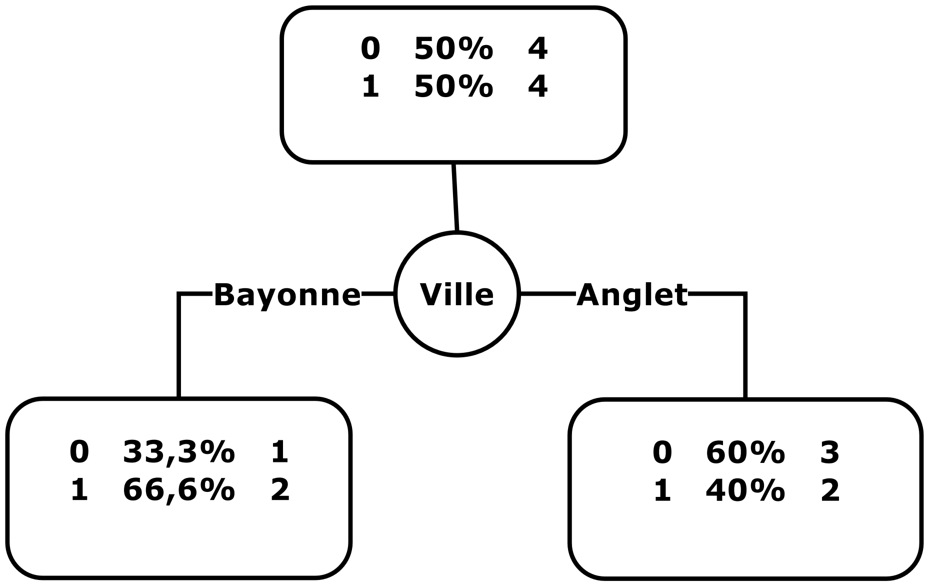
Pour le jeu de données ci-contre, quelle devrait être la première division ?

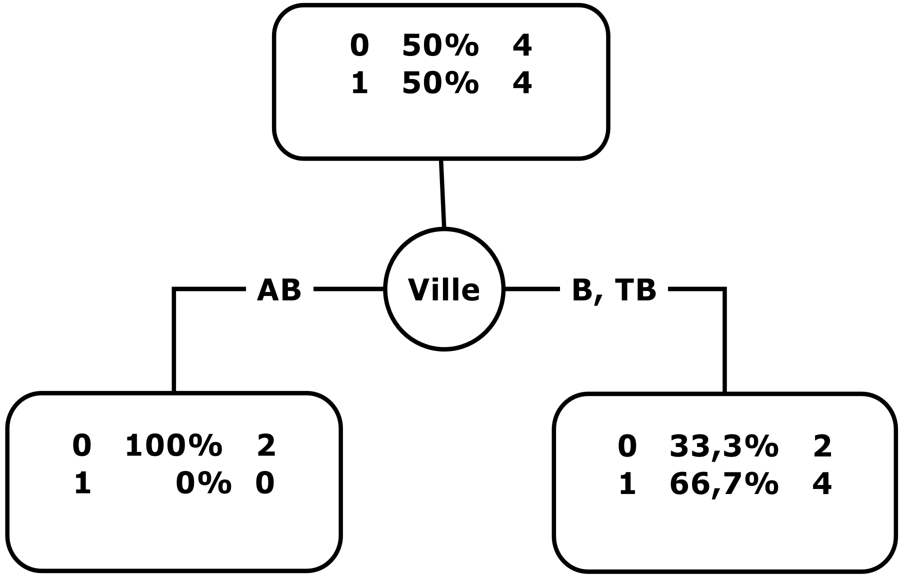
Ville	Avis	Cible
Bayonne	Très Bien	1
Anglet	Bien	1
Anglet	Assez Bien	0
Anglet	Assez Bien	0
Anglet	Bien	0
Bayonne	Très Bien	0
Anglet	Très Bien	1
Bayonne	Bien	1

Exercice

Trois divisions possibles :
Gain de pureté ?

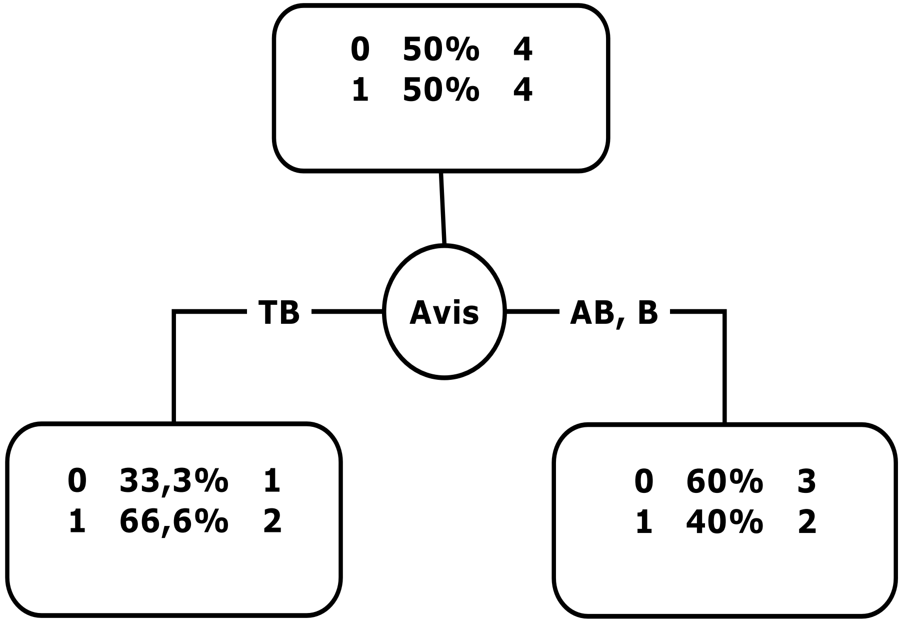






Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

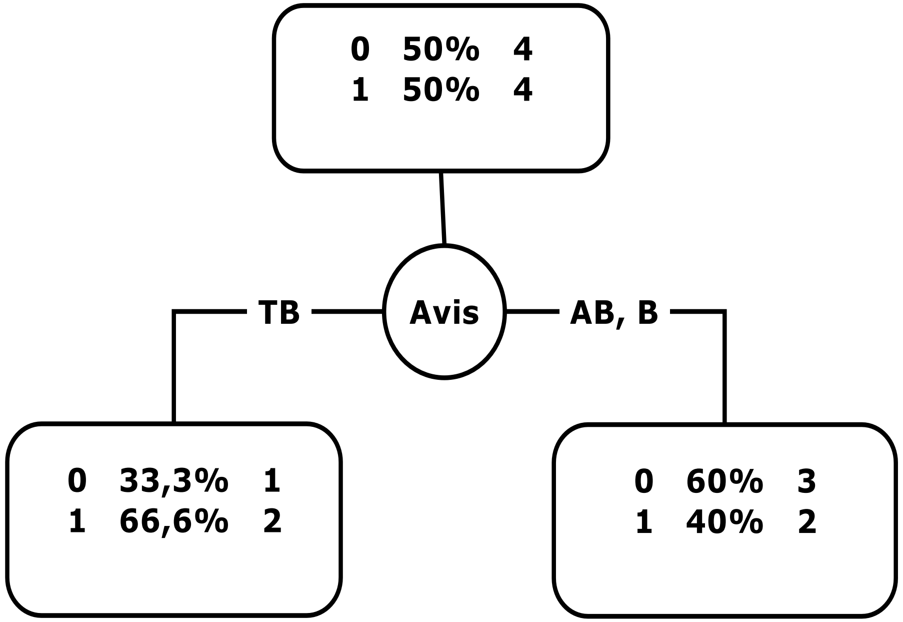
Exercice



$$Gini(\text{nœud père}) = ?$$

Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

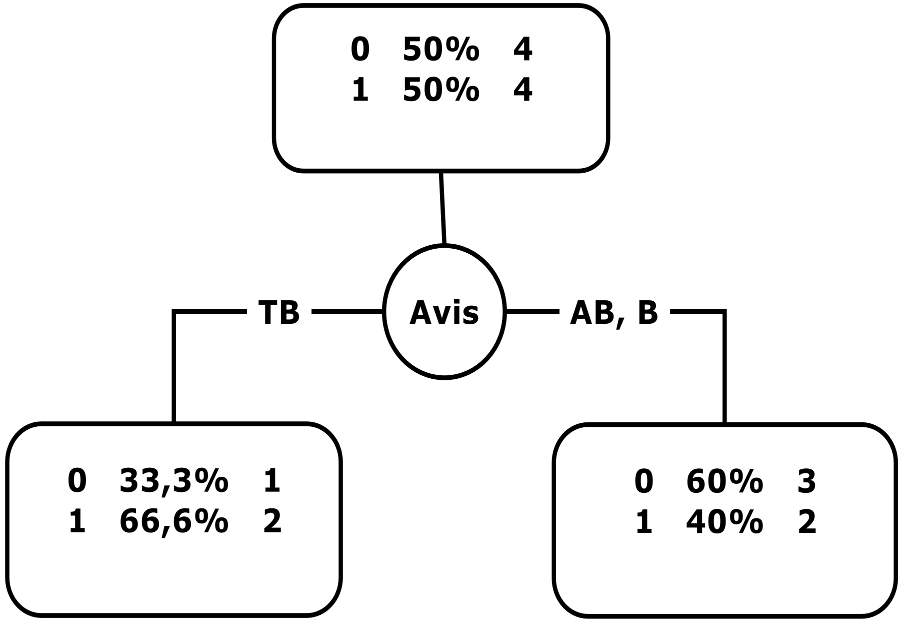
Exercice



$$Gini(\text{nœud père}) = 2f_0f_1 = 2 \times 0,5^2 = 0,5$$

Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

Exercice



$$Gini(\text{nœud père}) = 2f_0f_1 = 2 \times 0,5^2 = 0,5$$

$$G_G = Gini(\text{nœud gauche}) = 2f_0f_1 = 2 \times \frac{1}{3} \times \frac{2}{3}$$

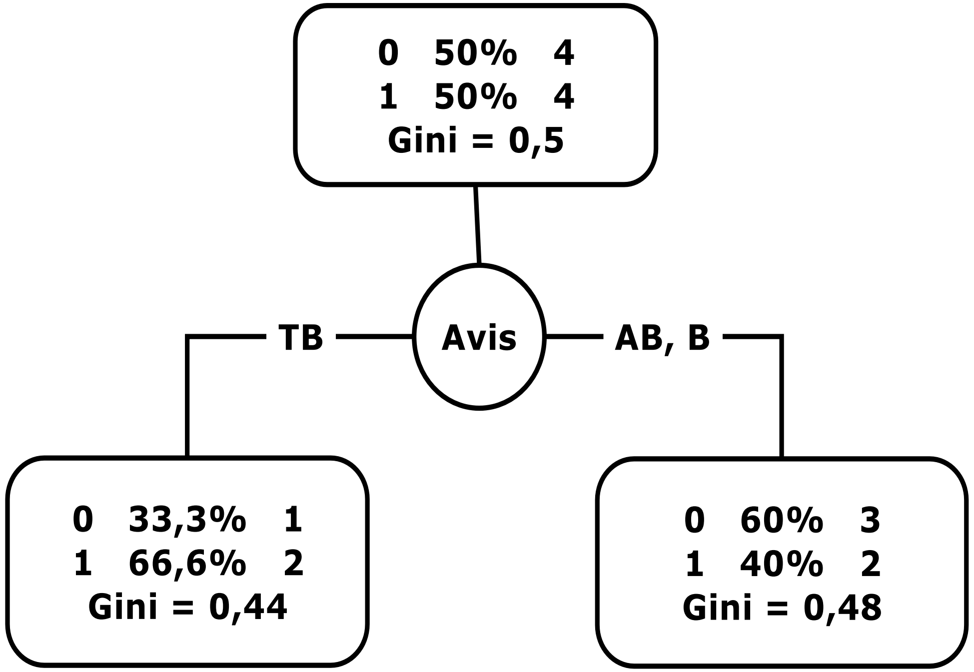
$$G_G = \frac{4}{9} = 0,44$$

$$G_D = Gini(\text{nœud droite}) = 2f_0f_1 = 2 \times \frac{3}{5} \times \frac{2}{5}$$

$$G_D = \frac{12}{25} = 0,48$$

Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

Exercice



$$Gini(\text{nœud père}) = 2f_0f_1 = 2 \times 0,5^2 = 0,5$$

$$G_G = Gini(\text{nœud gauche}) = 2f_0f_1 = 2 \times \frac{1}{3} \times \frac{2}{3}$$

$$G_G = \frac{4}{9} = 0,44$$

$$G_D = Gini(\text{nœud droite}) = 2f_0f_1 = 2 \times \frac{3}{5} \times \frac{2}{5}$$

$$G_D = \frac{12}{25} = 0,48$$

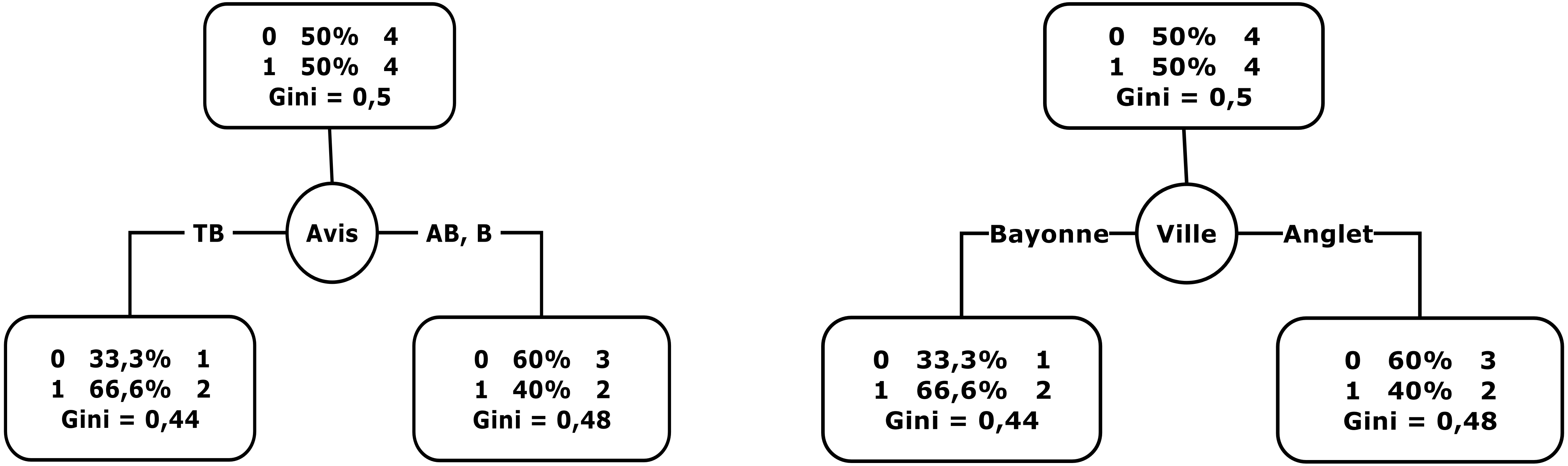
moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

$$G_G \times \frac{3}{8} + G_D \times \frac{5}{8} = \frac{4}{9} \times \frac{3}{8} + \frac{12}{25} \times \frac{5}{8} = \frac{21}{45} = 0{,}47$$

Gain de pureté : $0,5 - 0,47 = 0,03$

Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

Exercice

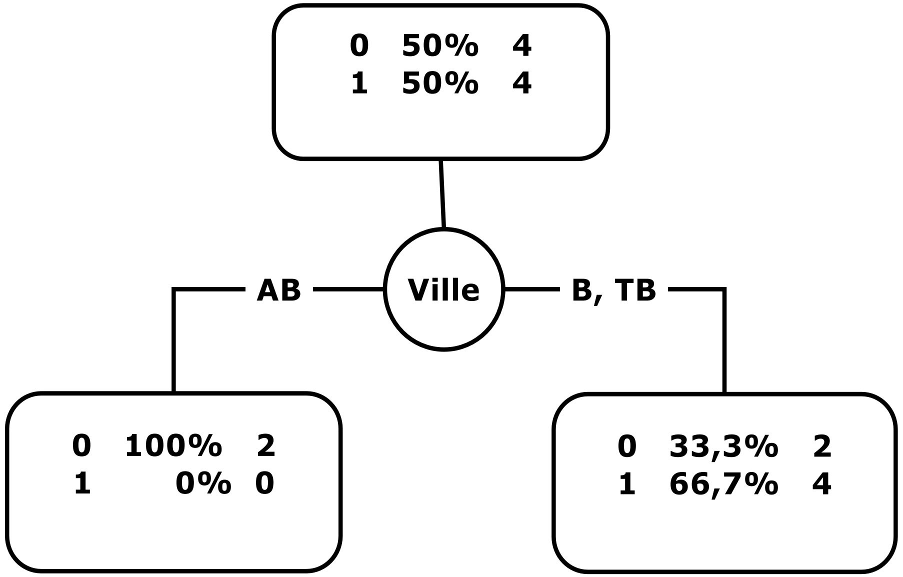


Mêmes indices de Gini sur les feuilles $\Rightarrow$ mêmes gains de pureté : 0,03

Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

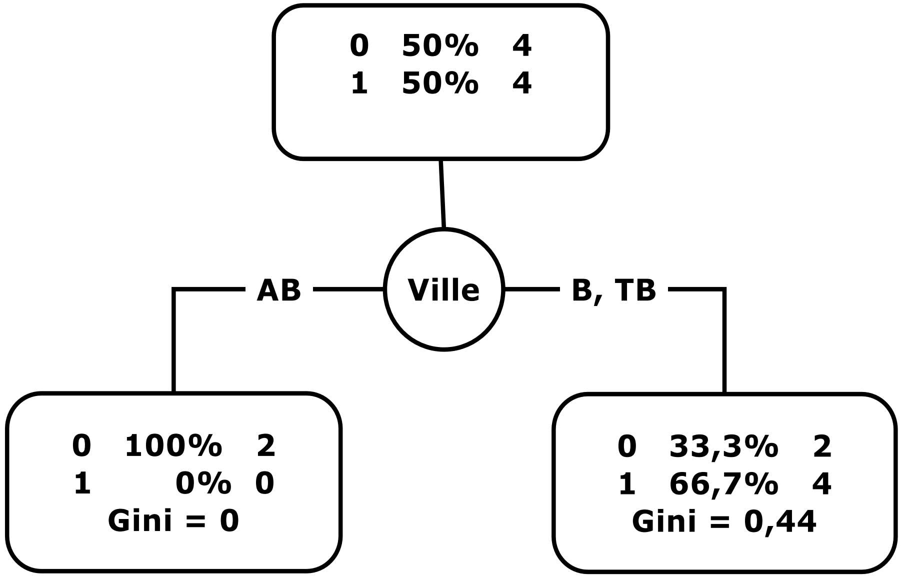
Exercice

$$Gini(\text{nœud père}) = 2f_0 f_1 = 2 \times 0,5^2 = 0,5$$



Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

Exercice



$$Gini(\text{nœud père}) = 2f_0f_1 = 2 \times 0,5^2 = 0,5$$

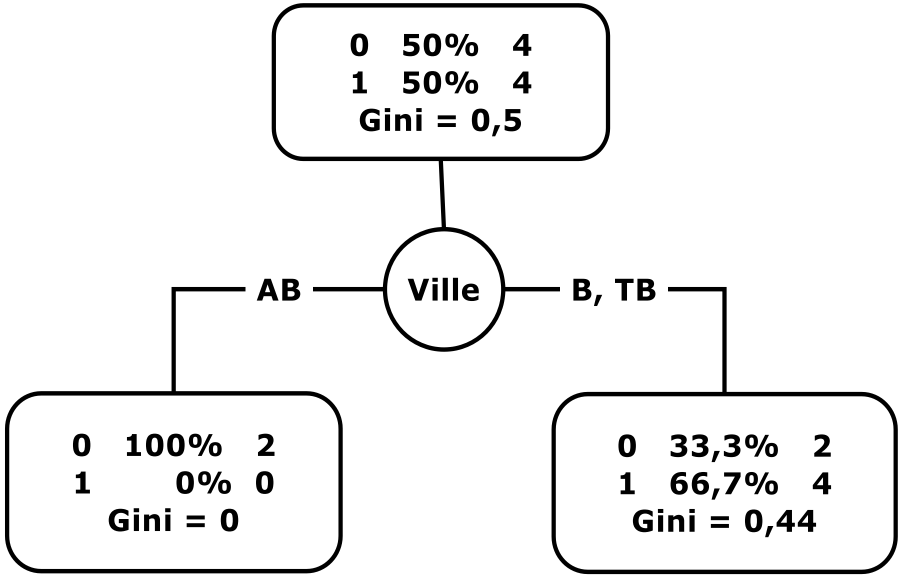
$$G_G = Gini(\text{nœud gauche}) = 2f_0f_1 = 2 \times 0 \times 1 = 0$$

$$G_G = 0$$

$$G_D = Gini(\text{nœud droite}) = 2f_0f_1 = 2 \times \frac{2}{6} \times \frac{4}{6}$$

$$G_D = \frac{4}{9} = 0,44$$

Exercice



Gain=Gini(nœud père) – moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

$$Gini(\text{nœud père}) = 2f_0f_1 = 2 \times 0,5^2 = 0,5$$

$$G_G = Gini(\text{nœud gauche}) = 2f_0f_1 = 2 \times 0 \times 1 = 0$$

$$G_G = 0$$

$$G_D = Gini(\text{nœud droite}) = 2f_0f_1 = 2 \times \frac{2}{6} \times \frac{4}{6}$$

$$G_D = \frac{4}{9} = 0,44$$

moyenne des Gini(nœuds fils) pondérée par les poids des nœuds

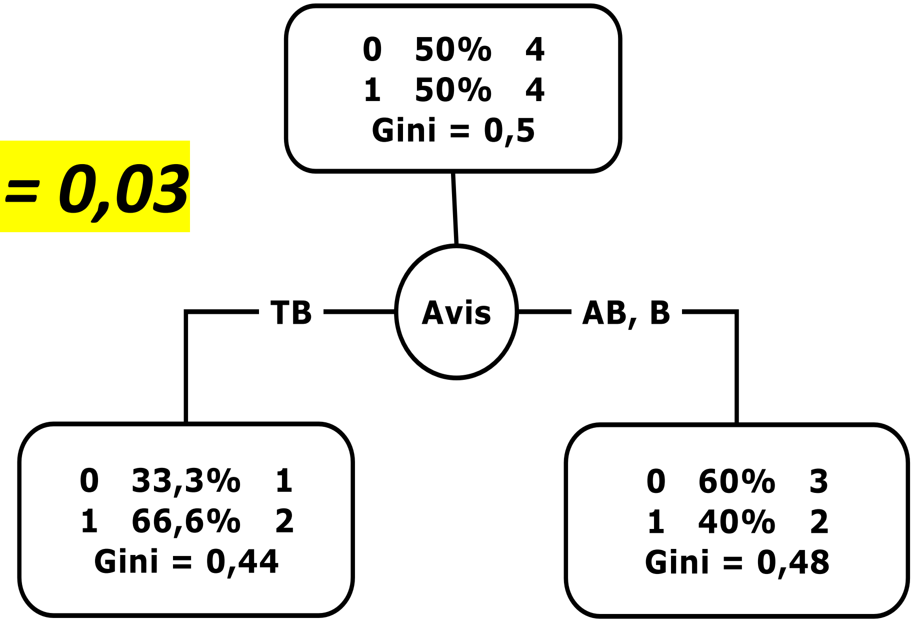
$$G_G \times \frac{2}{8} + G_D \times \frac{6}{8} = 0 \times \frac{2}{8} + \frac{4}{9} \times \frac{6}{8} = \frac{1}{3} = 0{,}33$$

Gain de pureté : $0,5 - 0,33 = 0,17$

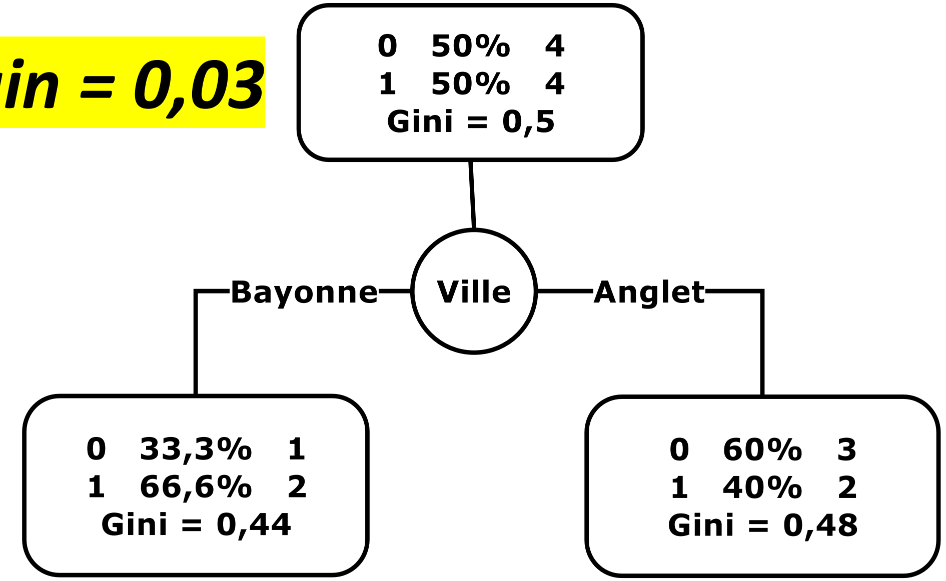
Exercice

Trois divisions possibles :
Gain de pureté ?

Gain = 0,03



Gain = 0,03



Gain = 0,17