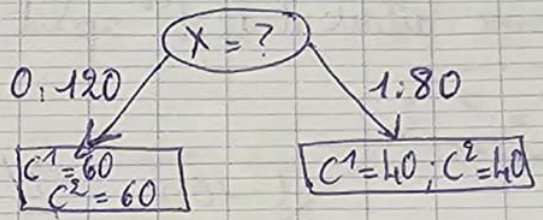
Nom : Sghiri Samir } TP Data Mining

Ex 7 :

a) X comme premier attribut pour le Split :



⇒ D'où le tableau de contingence associé à X :

Class \ X	0	1	Total
C^1	60	40	100
C^2	60	40	100
Total	120	80	200

→ Calculons d'abord l'erreur avant Split :

$$Erreur_{origine} = 1 - \max\left\{P(C^1), P(C^2)\right\}$$
$$= 1 - \max\left\{\frac{100}{200}, \frac{100}{200}\right\}$$
$$= 0,5$$

Notons bien que : $\Delta(X) = Erreur_{origine} - Erreur(X)|_{split}$

→ Calculons donc $Erreur(X)|_{split}$:

$$Erreur(X)|_{split} = \frac{120}{200} \cdot Erreur(X=0) + \frac{80}{200} \cdot Erreur(X=1)$$

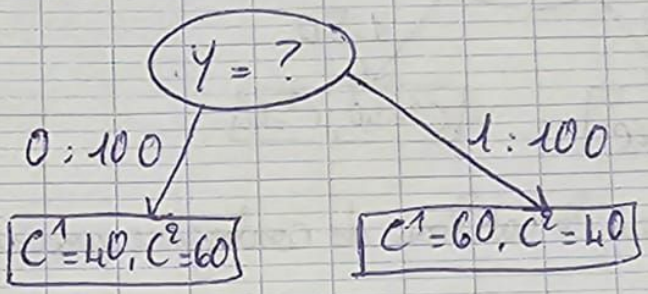
$$Erreur(X=0) = 1 - \max\left\{\frac{60}{120}, \frac{60}{120}\right\} = 0,5$$

$$Erreur(X=1) = 1 - \max\left\{\frac{40}{80}, \frac{40}{80}\right\} = 0,5$$

Donc : $Erreur(X)|_{split} = (0,6 * 0,5) + (0,4 . 0,5)$
$= 0,5$

D'où $\boxed{\Delta(X) = 0}$ $\Rightarrow$ Aucune amélioration

$\rightarrow$ Y comme attribut du 1er split



D'où le tableau de contingence associé à Y :

Class \ Y	0	1	Total
C^1	40	60	100
C^2	60	40	100
Total	100	100	200

$Erreur(Y=0) = 1 - \max\{P(C^1), P(C^2)\}$
$= 1 - \max\left\{\frac{40}{100}, \frac{60}{100}\right\}$
$= 0,4$

$Erreur(Y=1) = 1 - \max\{P(C^1), P(C^2)\}$
$= 1 - \max\left\{\frac{60}{100}, \frac{40}{100}\right\}$
$= 0,4$

$Erreur(Y)|_{split} = \frac{100}{200} . Erreur(Y=0) + \frac{100}{200} . Erreur(Y=1)$
$= (0,5 . 0,4) + (0,5 . 0,4) = 0,4$

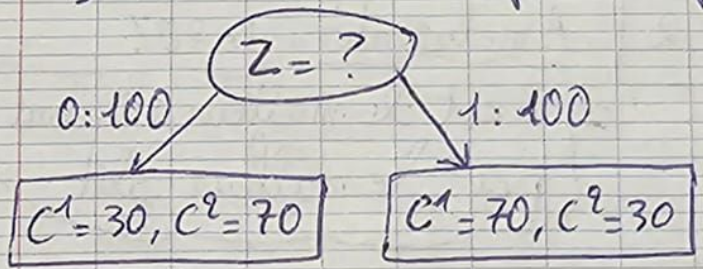
D'où : $\Delta(Y) = E_{meur\ origine} - E_{meur}(Y)|_{split}$

$= 0,5 - 0,4$

$\Rightarrow \boxed{\Delta(Y) = 0,1}$. Amélioration de 10% au niveau du taux d'erreur de Classification

$\simeq \boxed{Y \text{ est mieux que } X}$

→ Z comme attribut pour le premier split



D'où le tableau de contingence associé à l'attribut Z :

Class \ Z	0	1	Total
C^1	30	70	100
C^2	70	30	100
Total	100	100	200

$E_{meur}(Z=0) = 1 - \max\{P(C^1), P(C^2)\}$

$= 1 - \max\left\{\frac{30}{100}, \frac{70}{100}\right\}$

$= 0,3$

$E_{meur}(Z=1) = 1 - \max\{P(C^1), P(C^2)\}$

$= 1 - \max\left\{\frac{70}{100}, \frac{30}{100}\right\}$

$= \frac{30}{100} = 0,3$

$$E_{meur}(Z))_{split} = \frac{100}{200} \cdot E_{meur}(Z=0) + \frac{100}{200} \cdot E_{meur}(Z=1)$$

$$= (0,5 \cdot 0,3) + (0,5 \cdot 0,3)$$
$$= 0,3$$

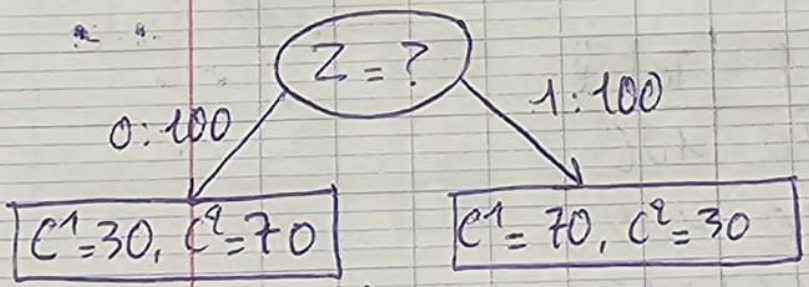
D'où :

$$\Delta(Z) = E_{meur\ origine} - E_{meur}(Z))_{split}$$
$$= 0,5 - 0,3$$

$\Rightarrow$ **$\Delta(Z) = 0,2$**, Amélioration de 20%

$\approx$ **Z est le meilleur attribut pour le meilleur split**

Donc, l'arbre de décision après le 1er split est :



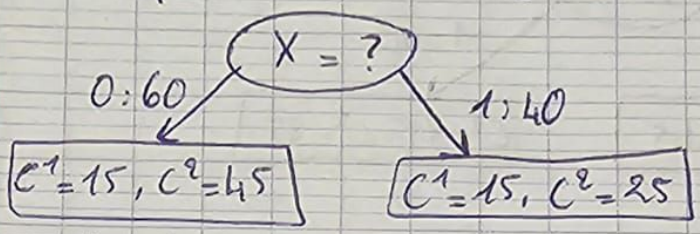
$\Rightarrow$ 2ème Split

$\rightarrow$ le sous-Tableau obtenu sous l'hypothèse Z=0 :

X	Y	C^1	C^2
0	0	5	40
0	1	10	5
1	0	10	5
1	1	5	20

→ X pour le deuxième split, sachant $z=0$

~~$E_{meur}(X=0)|_{split\ z=0} = 1 - max\{\}$~~



$$E_{meur}(X=0)|_{z=0} = 1 - \max\left\{P(C^1), P(C^2)\right\}$$
$$= 1 - \max\left\{\frac{15}{60}, \frac{45}{60}\right\}$$
$$= \frac{1}{4}$$

$$E_{meur}(X=1)|_{z=0} = 1 - \max\left\{\frac{15}{40}, \frac{25}{40}\right\}$$
$$= \frac{3}{8}$$

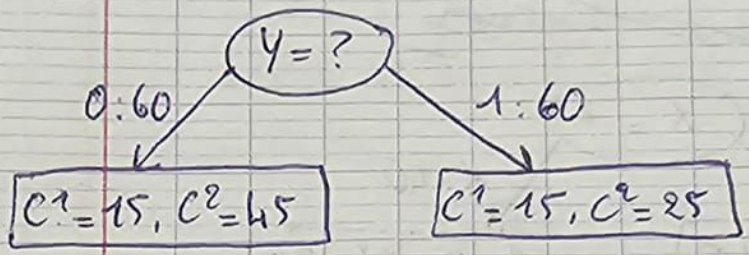
$$E_{meur}(X)|_{split}|_{z=0} = \frac{60}{100} \cdot E_{meur}(X=0)|_{z=0} + \frac{40}{100} \cdot E_{meur}(X=1)|_{z=0}$$
$$= 0,6 \cdot \frac{2}{8} + 0,4 \cdot \frac{3}{8}$$
$$= 0,3$$

$$E_{meur\ origine}|_{z=0} = E_{meur}(Z=0) = 0,3$$

D'où $\Delta(X)|_{z=0} = E_{meur}(Z=0) - E_{meur}(X)|_{split}|_{z=0}$
$= 0,3 - 0,3$

$\Rightarrow$ $\boxed{\Delta(X)|_{z=0} = 0}$: Aucune amélioration

→ Y pour le 2ème split sachant z=0



$$E_{meas}(Y=0)\big|_{z=0} = 1 - \max\{P(C^1), P(C^2)\}$$
$$= 1 - \max\left\{\frac{15}{60}, \frac{45}{60}\right\}$$
$$= \frac{1}{4}$$

$$E_{meas}(Y=1)\big|_{z=0} = 1 - \max\left\{\frac{15}{40}, \frac{25}{40}\right\}$$
$$= \frac{3}{8}$$

$$E_{meas}(Y)_{split}\big|_{z=0} = \left(\frac{60}{100}\cdot\frac{2}{8}\right) + \left(\frac{40}{100}\cdot\frac{3}{8}\right)$$
$$= 0,3$$

D'où

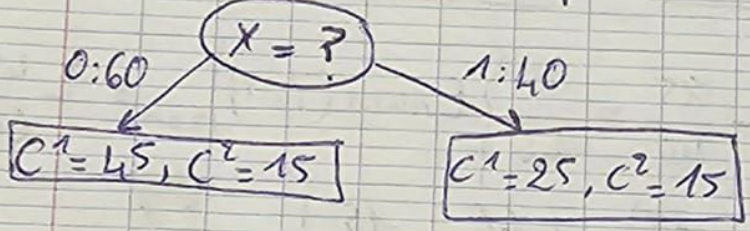
$$\Delta(Y)\big|_{z=0} = E_{meas}(z=0) - E_{meas}(Y)_{split}\big|_{z=0}$$
$$= 0,3 - 0,3$$

$$\Rightarrow \boxed{\Delta(Y)\big|_{z=0} = 0}$$: Aucune Amélioration

→ le sous-Tableau obtenu sous l'hypothèse $z=1$

X	Y	C^1	C^2
0	0	0	15
0	1	45	0
1	0	25	0
1	1	0	15

→ X pour le deuxième split, Sachant $z=1$



$$Emeur(X=0)\big|_{z=1} = 1 - \max\left\{P(C^1), P(C^2)\right\}$$
$$= 1 - \max\left\{\frac{45}{60}, \frac{15}{60}\right\}$$
$$= \frac{1}{4}$$

$$Emeur(X=1)\big|_{z=1} = 1 - \max\left\{\frac{25}{40}, \frac{15}{40}\right\}$$
$$= \frac{15}{40}$$

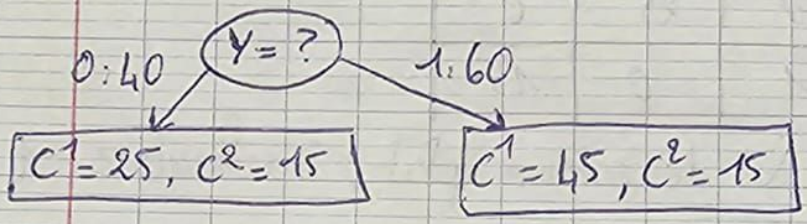
$$Emeur(X)\big|_{split}\big|_{z=1} = \frac{60}{100} \cdot Emeur(X=0)\big|_{z=1} + \frac{40}{100} \cdot Emeur(X=1)\big|_{z=1}$$
$$= \left(\frac{60}{100} \cdot \frac{1}{4}\right) + \left(\frac{40}{100} \cdot \frac{15}{40}\right)$$
$$= 0,3$$

D'où : $\Delta(X)|_{z=1} = Eneur(z=1) - Eneur(X)|_{split})_{z=1}$

$= 0,3 - 0,3$

$\Delta(X)|_{z=1} = 0$: Aucune Amélioration

→ Y pour le deuxième split, sachant $z=1$



$Eneur(Y=0)|_{z=1} = 1 - \frac{25}{40} = \frac{15}{40}$; $Eneur(Y=1)|_{z=1} = 1 - \frac{45}{60} = \frac{15}{60}$

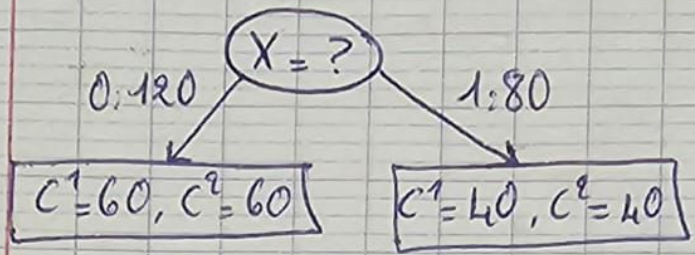
$Eneur(Y|_{split})|_{z=1} = \left(\frac{40}{100} \cdot \frac{15}{40}\right) + \left(\frac{60}{100} \cdot \frac{15}{60}\right)$

$= 0,3$

$\Delta(Y)|_{z=1} = 0,3 - 0,3 = 0$: Aucune Amélioration

Donc : le Taux d'eneur global de l'arbre induit est 30 %

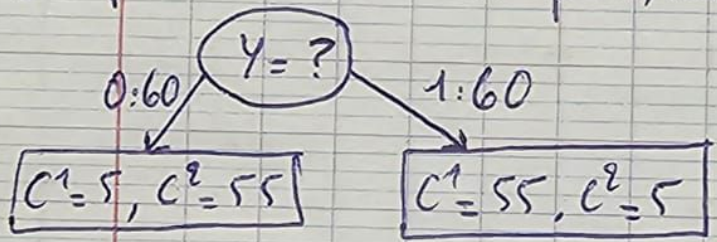
b) Utilisons X comme attribut pour le premier split
Donc, l'arbre de decision après le premier split est :



→ Passons on 2ème split
→ le sous-tableau obtenu sous l'hypothèse $X=0$

Y	Z	C^1	C^2
0	0	5	40
0	1	0	15
1	0	10	5
1	1	45	0

→ Y pour le deuxième split, sachant $X=0$



$$Erreur(X=0)\Big|_{X=0} = 1 - \max\left\{\frac{5}{60}, \frac{55}{60}\right\} = \frac{1}{12}$$

$$Erreur(X=1)\Big|_{X=0} = 1 - \max\left\{\frac{55}{60}, \frac{5}{60}\right\} = \frac{1}{12}$$

Page 10

$$E_{meur}(Y)\big|_{split}\big|_{x=0} = \frac{60}{120} \cdot E_{meur}(Y=0)\big|_{x=0} + \frac{60}{120} \cdot E_{meur}(Y=1)\big|_{x=0}$$

$$= \left(\frac{1}{2} \cdot \frac{1}{12}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) = \frac{1}{12} \simeq 0{,}08$$

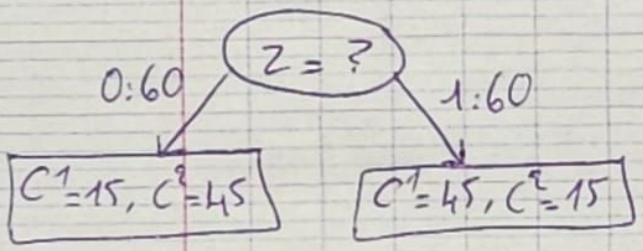
$$E_{meur\ origine}\big|_{x=0} = E_{meur}(X=0) = 0{,}5$$

$$\Rightarrow \Delta(Y)\big|_{x=0} = E_{meur}(X=0) - E_{meur}(Y)\big|_{split}\big|_{x=0}$$

$$= 0{,}5 - 0{,}08$$

$$\Rightarrow \boxed{\Delta(Y)\big|_{x=0} = 0{,}42}$$: Amélioration de 42 %.

$\longrightarrow$ Z pour le 2ème split, sachant $x=0$



$$E_{meur}(Z=0)\big|_{x=0} = 1 - \max\left\{\frac{15}{60}, \frac{45}{60}\right\}$$

$$= \frac{15}{60} = \frac{1}{4}$$

$$E_{meur}(Z=1)\big|_{x=0} = 1 - \max\left\{\frac{45}{60}, \frac{15}{60}\right\}$$

$$= \frac{1}{4}$$

$$E_{meur}(Z)_{split}\big|_{x=0} = \left(\frac{60}{120} \cdot \frac{1}{4}\right) + \left(\frac{60}{120} \cdot \frac{1}{4}\right)$$

$$= \frac{1}{4} = 0{,}25$$

Page 11

$\Delta(Z)\big|_{x=0} = 0,5 - 0,25 = 0,25$

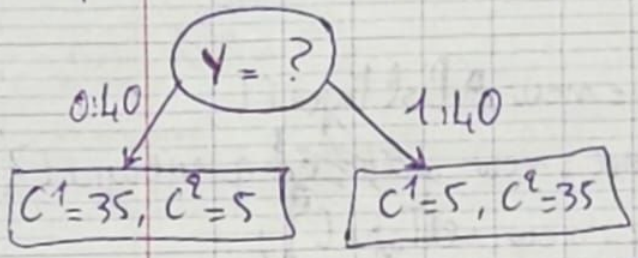
$\Rightarrow$ Amélioration de 25%

$\Rightarrow$ Donc Y est mieux suivant $X=0$

$\longrightarrow$ le sous-tableau obtenu sous l'hypothèse $X=1$

Y	Z	C^1	C^2
0	0	10	5
0	1	25	0
1	0	5	20
1	1	0	15

$\rightarrow$ Y pour la 2ème split, sachant $X=1$



$$\text{Emeur}(Y=0)\big|_{x=1} = 1 - \max\left\{\frac{35}{40}, \frac{5}{40}\right\} = \frac{1}{8}$$

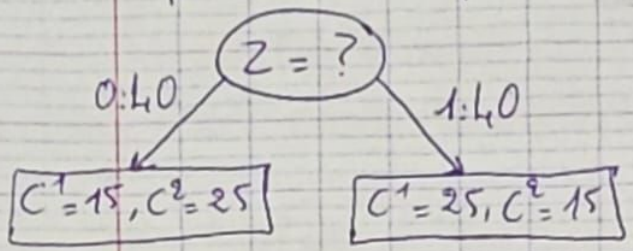
$$\text{Emeur}(Y=1)\big|_{x=1} = 1 - \max\left\{\frac{5}{40}, \frac{35}{40}\right\} = \frac{1}{8}$$

$$\text{Emeur}(Y)_{split}\big|_{x=1} = \left(\frac{40}{80}\cdot\frac{1}{8}\right) + \left(\frac{40}{80}\cdot\frac{1}{8}\right) = \frac{1}{8} \simeq 0,125$$

$\Delta(Y)\big|_{x=1} = 0,5 - 0,125 = 0,375 \Rightarrow$ Donc Amélioration de 37,5%.

Page 12

→ Z pour le 2ème split, sachant X = 1



$$\text{Erreur}(Z=0)\Big|_{X=1} = 1 - \max\left\{\frac{15}{40}, \frac{25}{40}\right\} = \frac{15}{40}$$

$$\text{Erreur}(Z=1)\Big|_{X=1} = 1 - \max\left\{\frac{25}{40}, \frac{15}{40}\right\} = \frac{15}{40}$$

$$\text{Erreur}(Z)_{\text{split}}\Big|_{X=1} = \frac{40}{80}\left(2 \cdot \frac{5}{40}\right)$$

$$= \frac{1}{2} \cdot 2 \cdot \frac{15}{40} = \frac{15}{40} = 0{,}375$$

$$\Delta(Z)\Big|_{X=1} = 0{,}5 - 0{,}375$$

$$= 0{,}125 \simeq 12{,}5\% \text{ d'amélioration}$$

Donc Y est mieux suivant X = 1

D'où le taux d'erreur de l'arbre induit est 37,5%

c) L'heuristique gourmande, bien qu'efficace pour la construction d'arbres de décision en minimisant le taux d'erreur à chaque étape, n'est pas exempte de limitations. Comme le montre le résultat obtenu dans la question (b). Le choix initial de l'attribut peut avoir un impact significatif sur le taux d'erreur global de l'arbre. En effet, si l'attribut sélectionné n'est pas optimal, cela peut entraîner une augmentation du taux d'erreur final.