

Le 1 Avril 2025

Introduction à l'arbre de Decision (DT)

1.) Retour sur la
classification kNN + Métriques. ✓

2.) Concept de Base du DT

3.) Critères utilisés pour le DT

- Entropie + Gain d'information (G.I).
- Critère GINI

4.) Progression - projet de Session

5.) Quiz formatif

classification : y est Discret

Les valeurs de $y \in \{l_1, l_2, \dots, l_n\}$

KNN : 1-) Valeur de K : les plus proches voisins

2-) Pas de training \Rightarrow Comparaison du point avec y inconnue à tous les points du dataset

3-) On utilise une distance pour mesurer l'écart entre les points.

4-) Sélection du label pour le y considéré sera faite selon

- Vote égalitaire

- Vote pondéré (selon la distance)

Inconvénient Temps de calcul !!

Tuning: 1) varier des valeurs autres pour le k.

2) evaluer d'autres distances.

R.L — RNL
Regularisation — Ridge
Lasso

classification :

KNN

Métrique

Accuracy

Precision

Sensibilité

F1-Score

AUC Area Under the Curve

objectif: Appris & Développer un modèle.

—

Classification

Decision Tree.

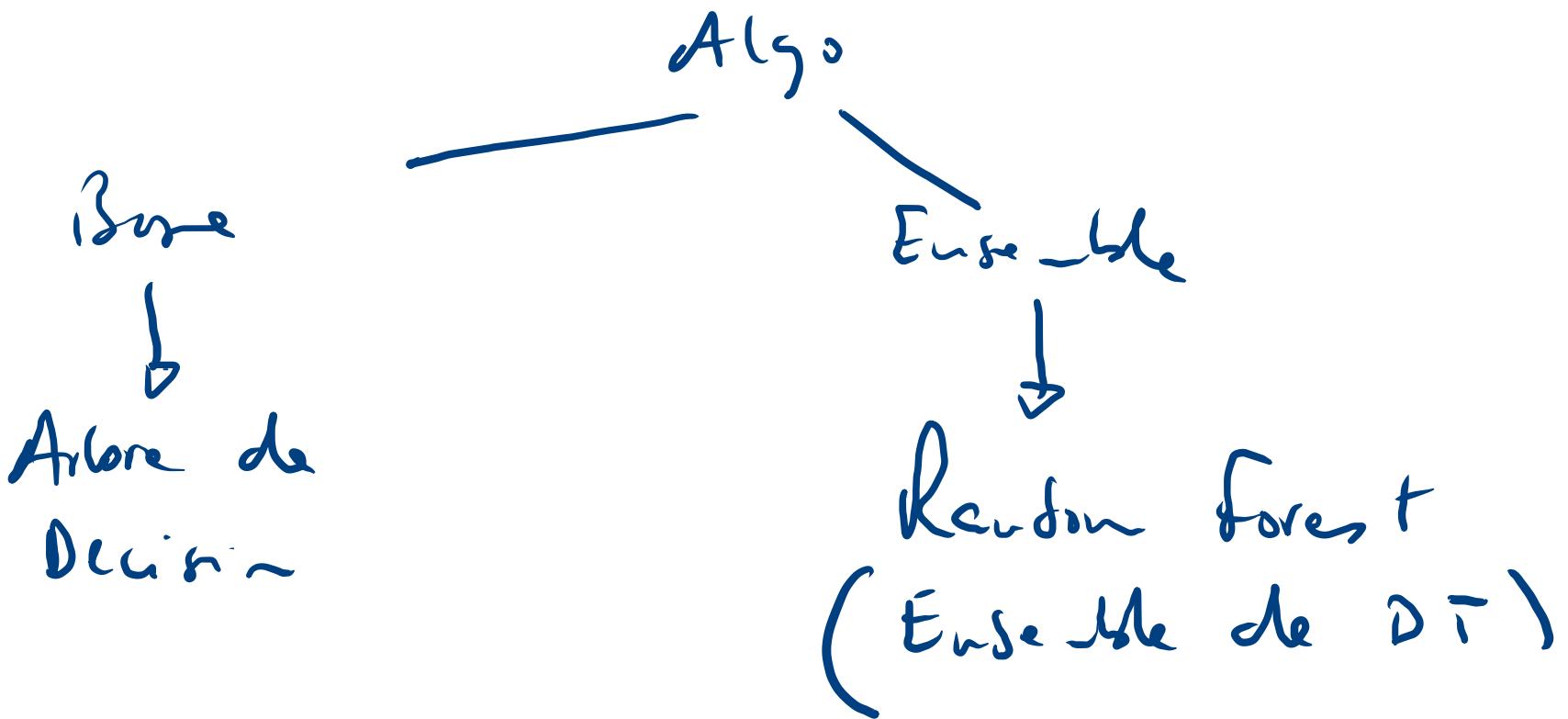
logistic Regres

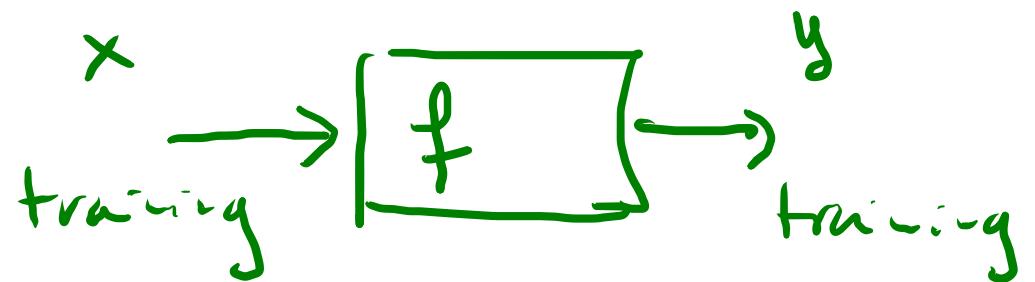
= Naive Bayes.

Approches d'ensem

DT → Random
Forest

Concept du Decision Tree.





KNN

$\text{Mod. } f \text{ fit}(x_{\text{train}}, y_{\text{train}})$

$\text{Mod}(\text{KNN}) \rightarrow \text{Metr}(\text{KNN})$

Accuracy:

$x_{\text{test}} \rightarrow \text{Mod}(\text{KNN}) \rightarrow y_{\text{pred}}$

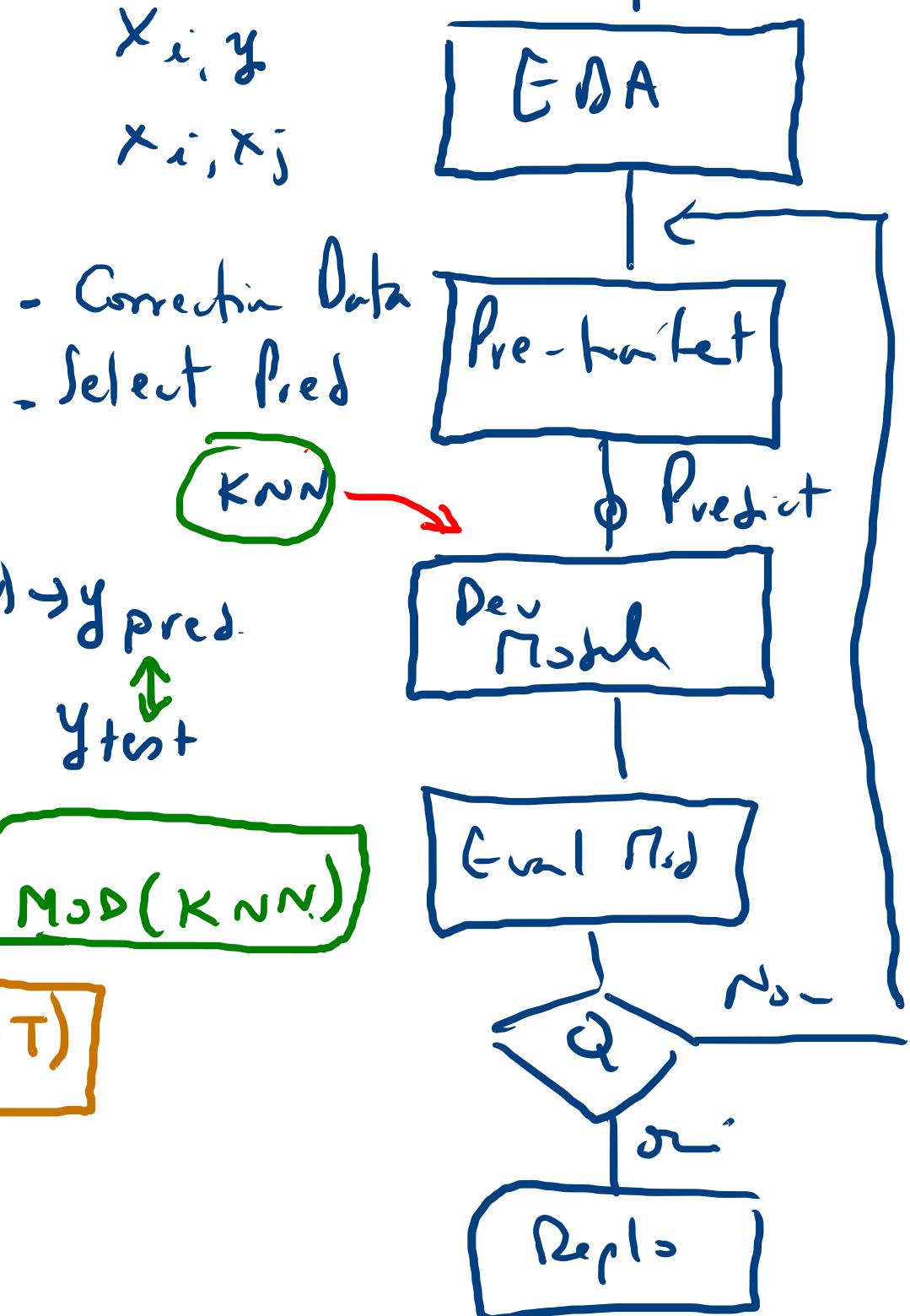
$\text{Mod}(\text{KNN})$

$\text{Acc} > 50\%$

\Rightarrow Plusieurs iterations

Accuracy: 82% → $\text{Mod}(\text{KNN})$

↳ DT → Accuracy: 92% → $\text{Mod}(\text{DT})$



1980

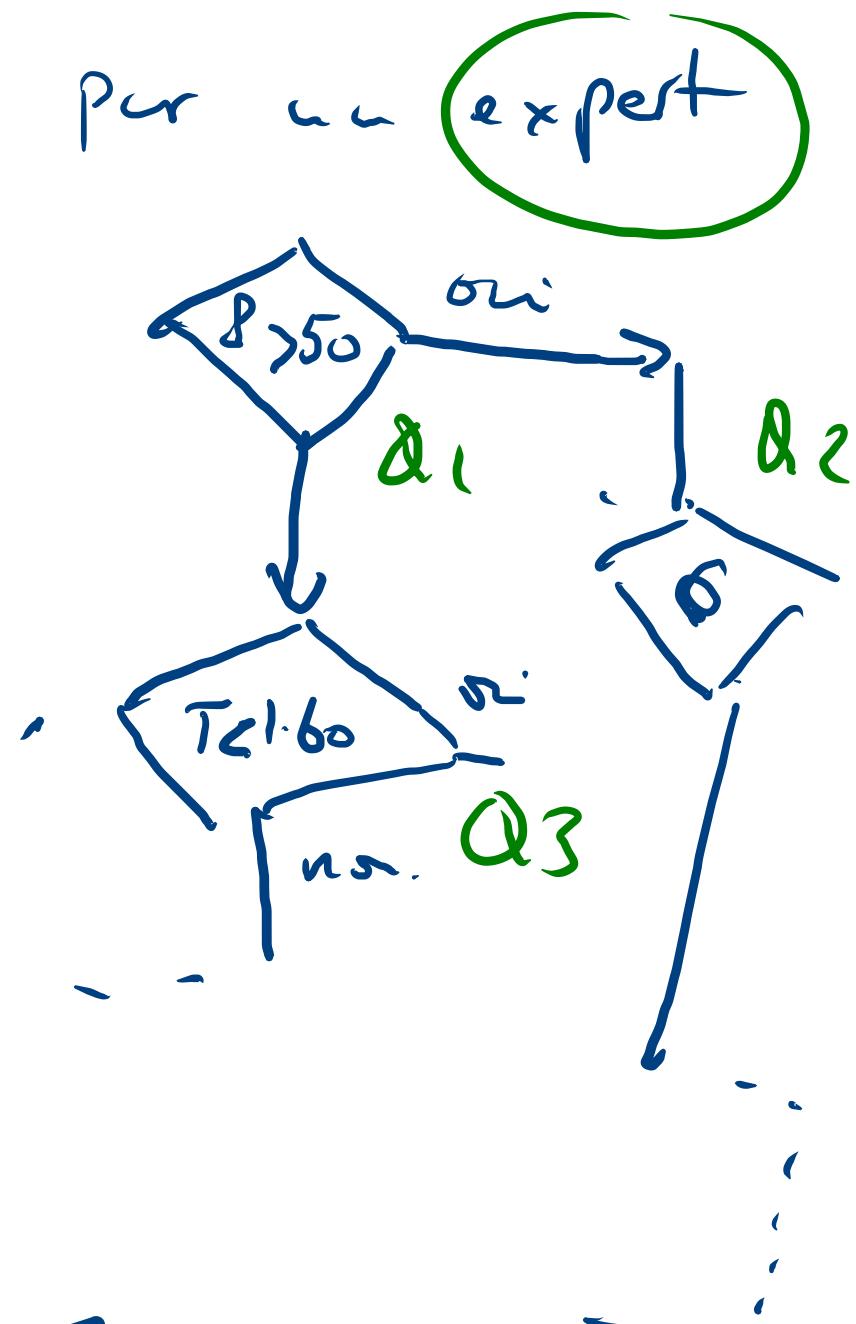
Système Expert

Série de Questions : Développée par un expert du Domaine

Expert : Sélectionne les questions faisant partie du système.

But : Identifie la réponse qui correspond à la meilleure classe.

- Dépend de l'expert du Domaine ; questions sont subjectives
- Absence de Data d'entraînement

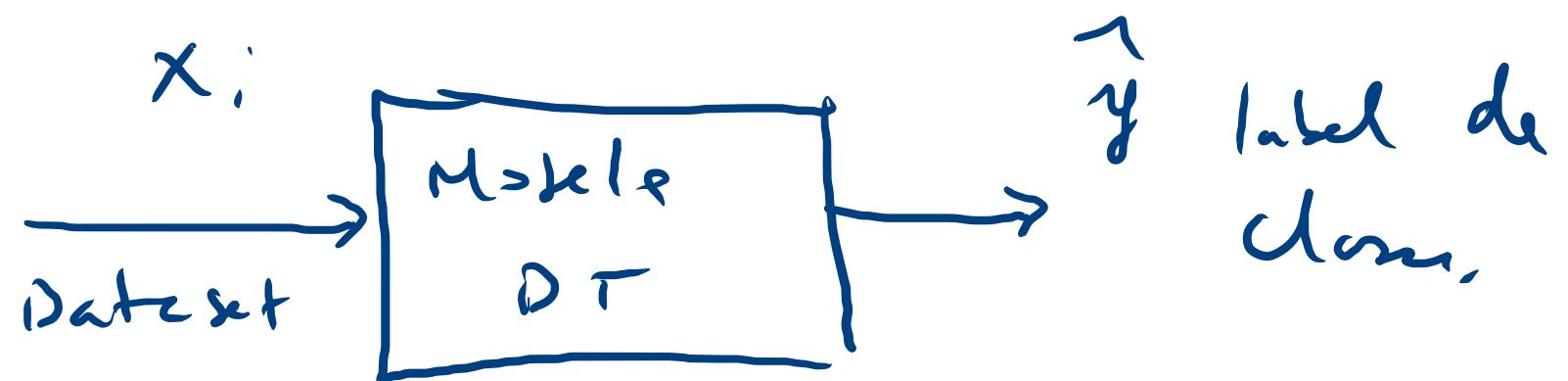


R_P
Etudiant
est n_o g_e

R_GMN.
Etudiant
est un
genie.

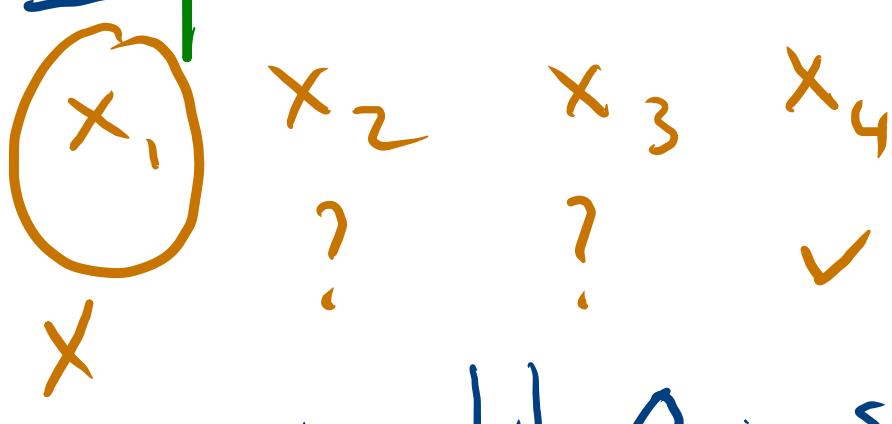
Sachant le data
de training \Rightarrow Approche objective.
 \Rightarrow Questions et le
sequencement.

Objectif: Etablir la séorie de questions (Decision)
basée sur le data présent au
niveau du Dataset (X_i).



\Rightarrow Automatise
- le choix des questions
- la sequence

nom	porte Lurette	Explor	Note intra	y
-	-	-	-	Echec
-	-	-	-	Succes
-	-	-	-	Succes
-	-	-	-	Echec .



Q1: si note intra

$$x_4 > 60$$

sinon

y: Echec

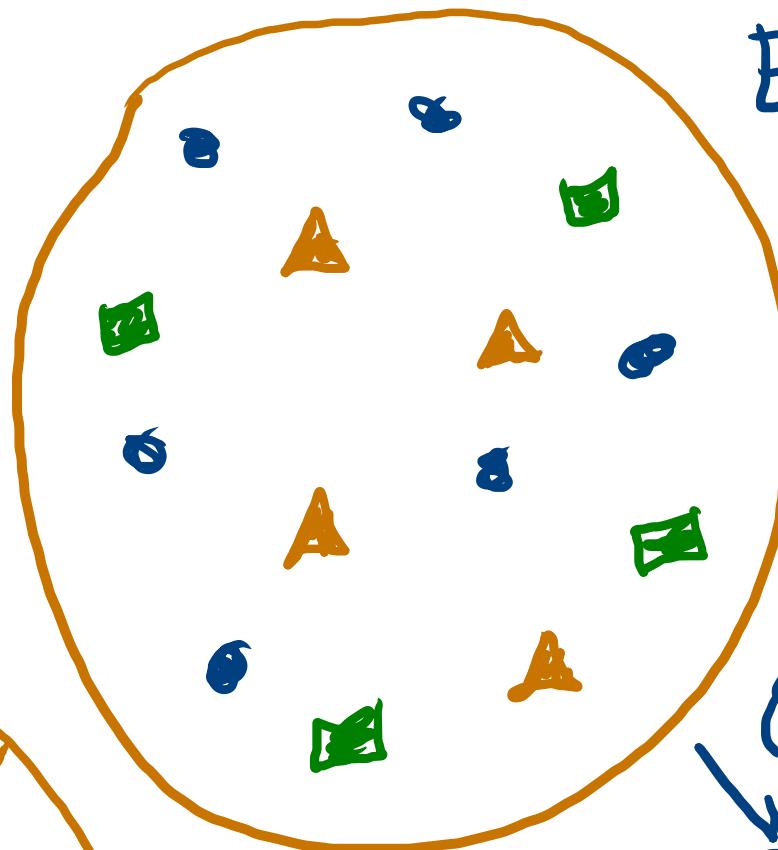
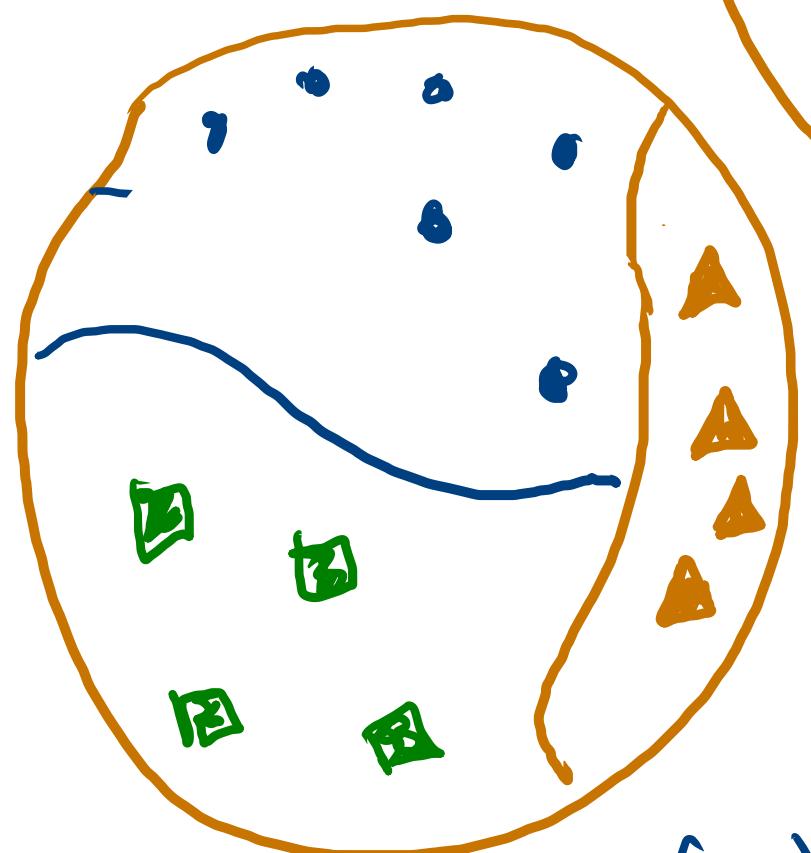
Q1: si en Explor

$$x_3$$

sinon

y: Succes

Mélange : Hétérogène



$Q(?)$

$Q(?)$ ($\times 4$ M_K I_T)

• l_1
△ l_2
□ l_3

E_2

- 1) choix des Questions. } Mense Entropie / G-I
2) Ordre des Questions.
3) Comment mesurer Hétérogénéité / Homogénéité.



Mense du Desordre



ENTROPIE

À chaque niveau de question

⇒ On mesure Entropie.

⇒ on obtient $G_I = E_1 - E_2$

plus il est grand, plus la question est appropriée (pertinente)

$x_1 \quad x_2 \quad x_3$

Comment calculer
l'entropie ?

Entropie du système : E_S

$E(x_1)$

$$GI_1 := E_S - E(x_1)$$

le plus grand

$E(x_2)$

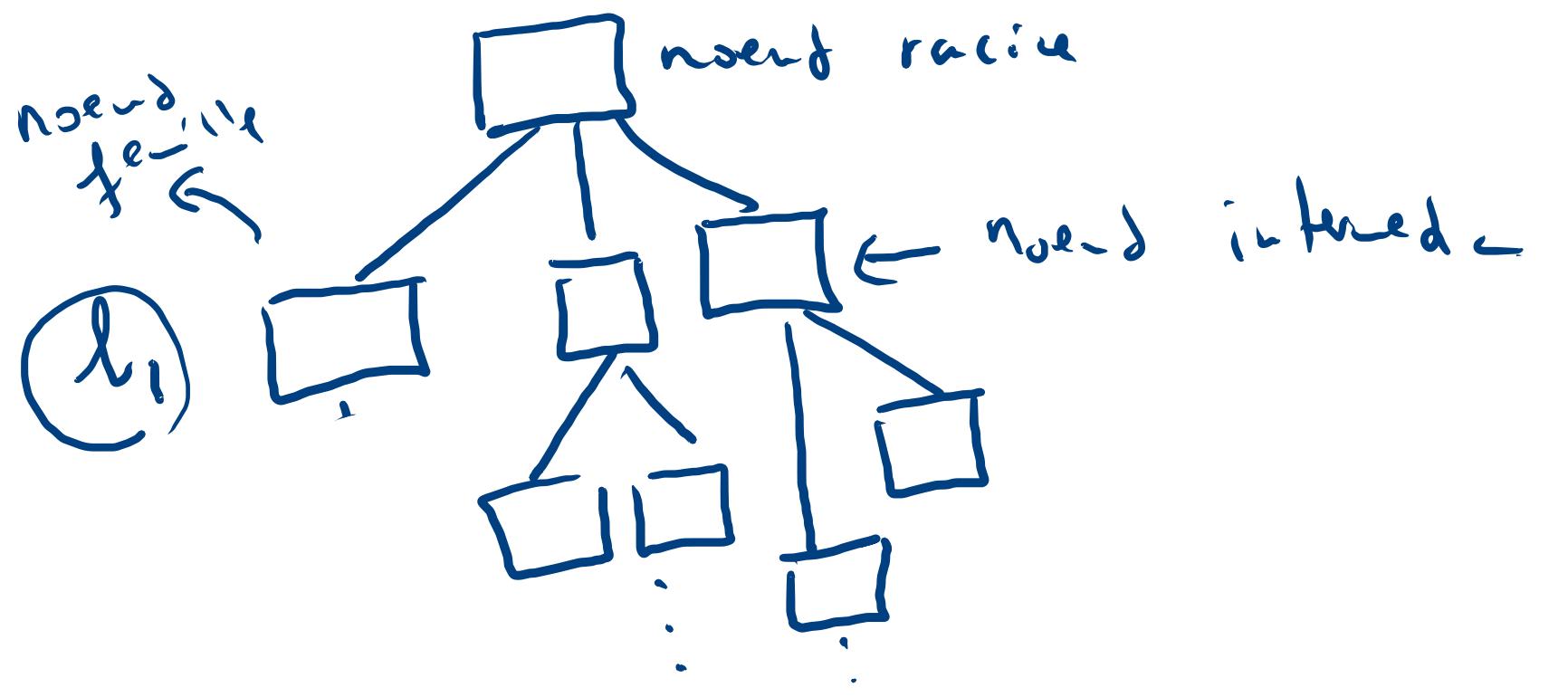
$$GI_2 := E_S - E(x_2)$$

Gain \Rightarrow

$E(x_3)$

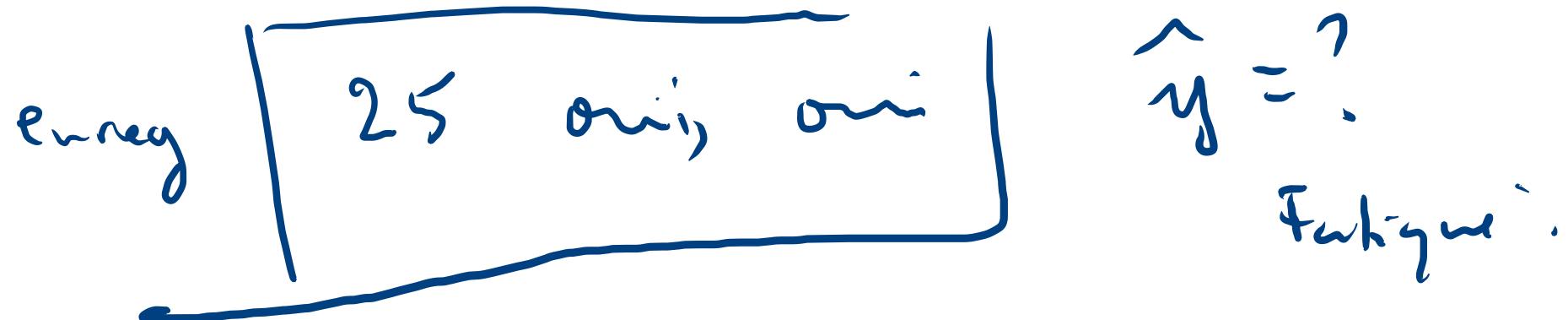
$$GI_3 := E_S - E(x_3)$$

le x à utiliser



x_1 age	x_2 Many Puls	x_3 Sport	y
30	oui	oui	Fatig
25	oui	non	Fatig
56	non	non	Fatigue
72	oui	non	Fon

y : Fatigué
En fon.



Ex : Déterminer le noeud racine candidat

Age, revenu, statut, credit

Le meilleur sera celui qui a le plus grand gain d'information.

Etape 1 : Calculer l'entropie du système initial

$$H(y) = - \sum_{y \in E} p(y) \log_2 p(y)$$

$E: \{\text{oui}, \text{non}\}$

nbre Labels: 2

Nbre Entreg = 14.

$$p(y=\text{oui}) = \frac{9}{14}$$

$$p(y=\text{non}) = \frac{5}{14}$$

$$\begin{aligned} E_{\text{initial}} &= - \frac{9}{14} \log \frac{9}{14} - \frac{5}{14} \log \frac{5}{14} \\ &= 0.94 \end{aligned}$$

Etape 2: Calcul du E^I pour $X = \text{Age}$

Pour chaque occurrence de $X = \text{age}$ on fait le calcul pour y .
af 0.6

catégories de X

$$5 \quad \leq 30 (\leq 30) \quad E_A^1$$

$$5 \quad 31-40 \quad E_A^2$$

> 30

$$4 \quad > 40 \quad E_A^3$$

$$E_A^1 \quad P(y=\text{oui} | \text{age} \leq 30) = \frac{2}{5} \Rightarrow E_A^1 = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5}$$

$$P(y=\text{non} | \text{age} \leq 30) = \frac{3}{5}$$

$$E_A^2 \quad P(y=\text{oui} | \text{age } 31-40) = \frac{5}{5} = 1$$

$$P(y=\text{non} | \text{age } 31-40) = \frac{0}{5} = 0$$

$$E_A^3 \quad P(y=\text{oui} | \text{age} > 40) = \frac{2}{4} = \frac{1}{2}$$

$$P(y=\text{non} | \text{age} > 40) = \frac{2}{4} = \frac{1}{2}$$

$$E_{\text{Age}} = \frac{5}{14} E_A^1 + \frac{5}{14} E_A^2 + \frac{4}{14} E_A^3$$

$$= \frac{5}{14} \times 0.97 + \frac{5}{14} \times 0 + \frac{4}{14} \times 1 = 0.63$$

$$E_A^1 = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} \approx 0.97$$

$$E_A^2 = -1 \log 1 - 0 \log 0 = 0$$

$$E_A^3 = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$$

$$GI_{Age} = E_{Initial} - E_{Age} = 0.94 - 0.63 = \boxed{0.31}$$

_____.

E type 3 : E_{start}

$$\textcircled{7} \quad \frac{E^1_{\text{start}}}{E_{\text{start}}} =$$

$$p(y_{\text{non}} | \text{start} = \text{ori}) = \frac{6}{7}$$

$$p(y_{\text{non}} | \text{start} = \text{non}) = \frac{1}{7}$$

$$E_{\text{st}}^1 = -\frac{6}{7} \log \frac{6}{7} - \frac{1}{7} \log \frac{1}{7} \\ = 0.59$$

$$\textcircled{7} \quad \frac{E^2_{\text{start}}}{E_{\text{start}}} =$$

$$p(y_{\text{non}} | \text{start} = \text{non}) = \frac{3}{7} \\ p(y_{\text{non}} | \text{start} = \text{ori}) = \frac{4}{7}$$

$$E_{\text{st}}^2 = -\frac{3}{7} \log \frac{3}{7} \\ - \frac{4}{7} \log \frac{4}{7} \\ = 0.97$$

$$E_{\text{start}} = \frac{7}{14} E_{\text{start}}^1 + \frac{7}{14} E_{\text{start}}^2$$

$$= \frac{7}{14} \times 0.59 + \frac{7}{14} \times 0.97 = 0.79$$

$$G I_{\text{start}} = E_{\text{initial}} - E_{\text{start}} = 0.94 - 0.79 = \boxed{0.15}$$

$$GI_{\text{Revenue}} = 0.029$$



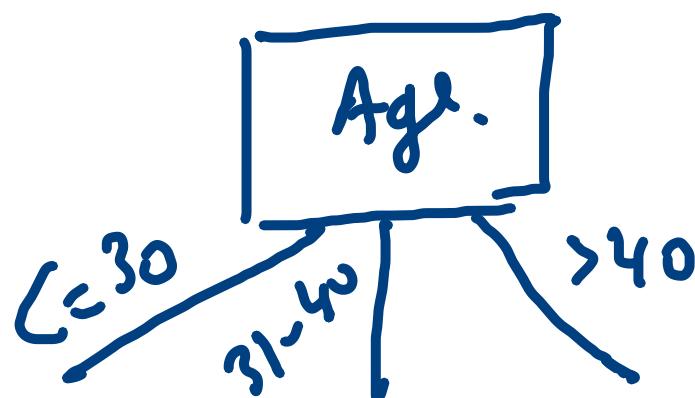
$$GI_{\text{Credit}} = 0.048$$

$$GI_{\text{Age}} = 0.31$$

No end Racine



AGE



Pratique Notebook DT;

- Reprend Notebook KNN puis
recliner le modèle en se basant sur le
decision tree

NB

- ouverture du Dataset. *breast-data.*
- EDA *corr.*
- creation du modèle. *split train-test* *instantiate model.* *fit / predict*
- Calcul des metrics: *Accuracy | ...*
- exporter du modèle.