UMR 5205 CNRS

Apprentissage automatique

Haytham Elghazel

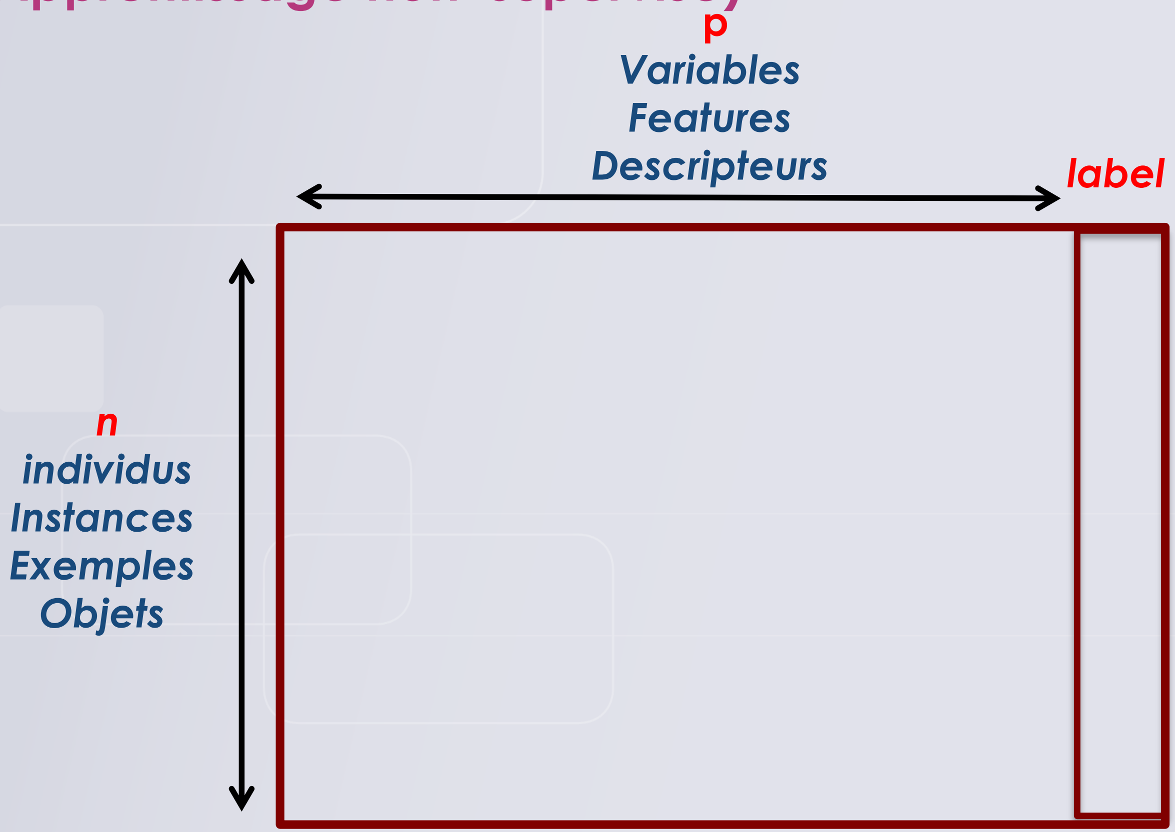
Laboratoire d'InfoRmatique en Image et Systèmes d'information

Pôle Data Science, Equipe DM2L

Contexte : Apprentissage automatique

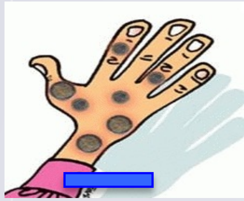
Consiste à inférer des connaissances sur les données

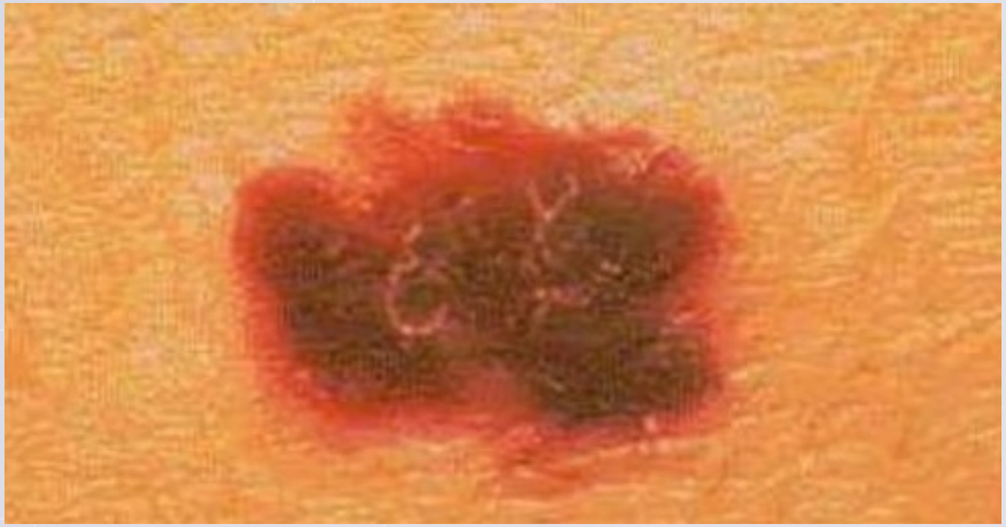
- Sur la seule base des échantillons d'apprentissage
- **Présence d'une cible (Apprentissage supervisé), Absence d'une cible (Apprentissage non supervisé)**



Apprentissage supervisé

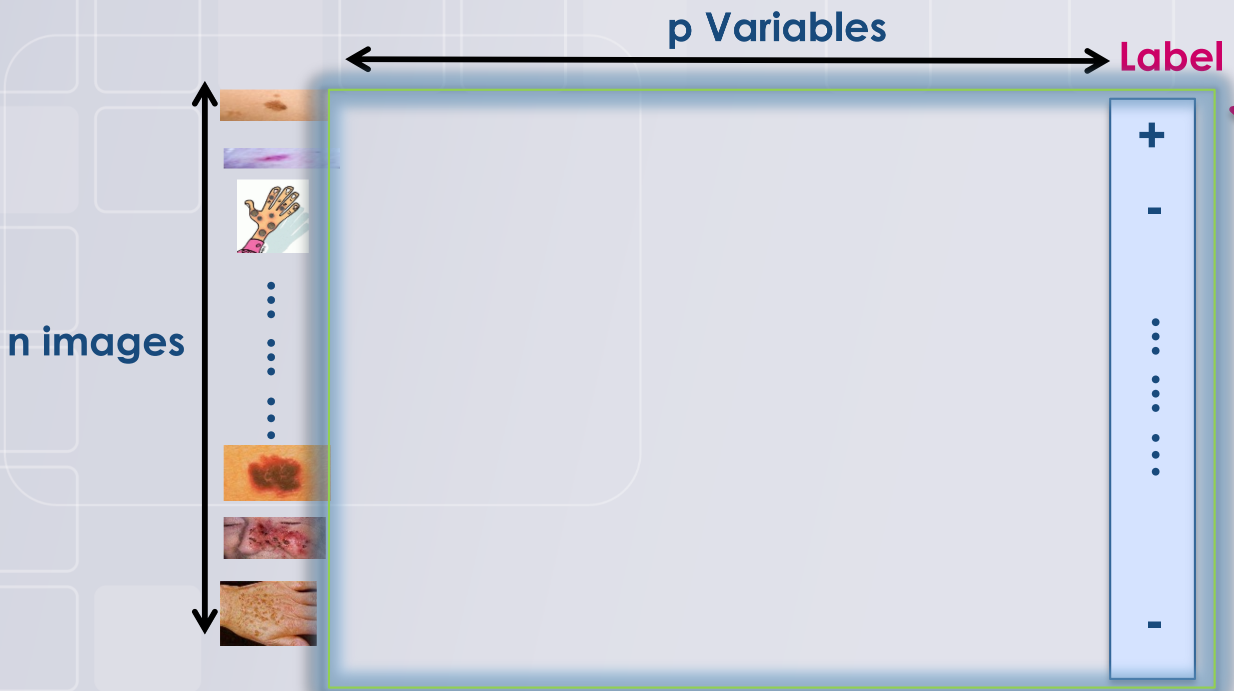






Est-ce qu’il s’agit d’un mélanome ?

Apprentissage supervisé



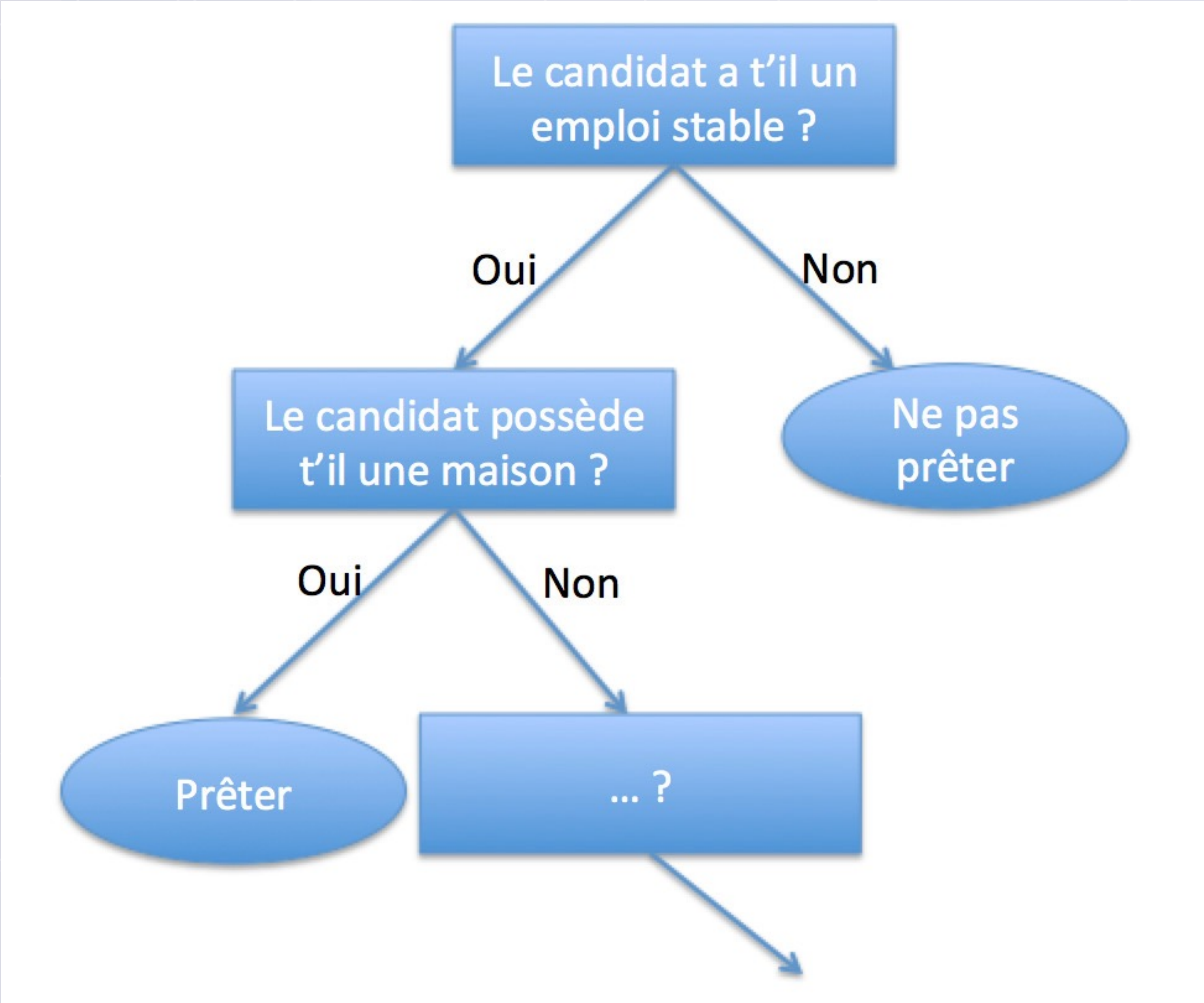
Données labélisées

- Généralement, une **méthode d'apprentissage Supervisé** est utilisée pour construire un **classifieur (ou modèle ou hypothèse)**, à partir d'une **base d'exemples étiquetés**, pour **prédire le label d'un nouvel individu qui arrive.**

Apprentissage supervisé

- **Apprentissage supervisé** : Produire automatiquement des règles à partir d'une base de données d’apprentissage étiquetées
- But: prédire la classe de nouvelles données observées
- Algorithmes : Arbres de décision, réseaux bayésiens, réseaux de neurones, k-plus proches voisins, etc…

Arbre de décision



Classement par les Arbres de Décision

- Processus récursif de division de l’espace des données en sous-régions de plus en plus pures en terme de classes.
- Décomposition d’un problème de classification en une suite de tests (imbriqués) portant sur une variable (parallèle aux axes) ou une combinaison linéaire de plusieurs variables (oblique).
- Les tuples sont placés dans la classe associée à la sous-région qu’ils vérifient
- Différentes approches en fonction de la façon dont l’arbre est construit

Arbre de décision : Exemple

- On souhaite construire un arbre de décision qui soit capable de déterminer si un client consultera son compte par Internet en fonction des **4 attributs** : **M** Moyenne du solde de comptes, **A** Age, **R** Résidence et **E** Niveau d'études.

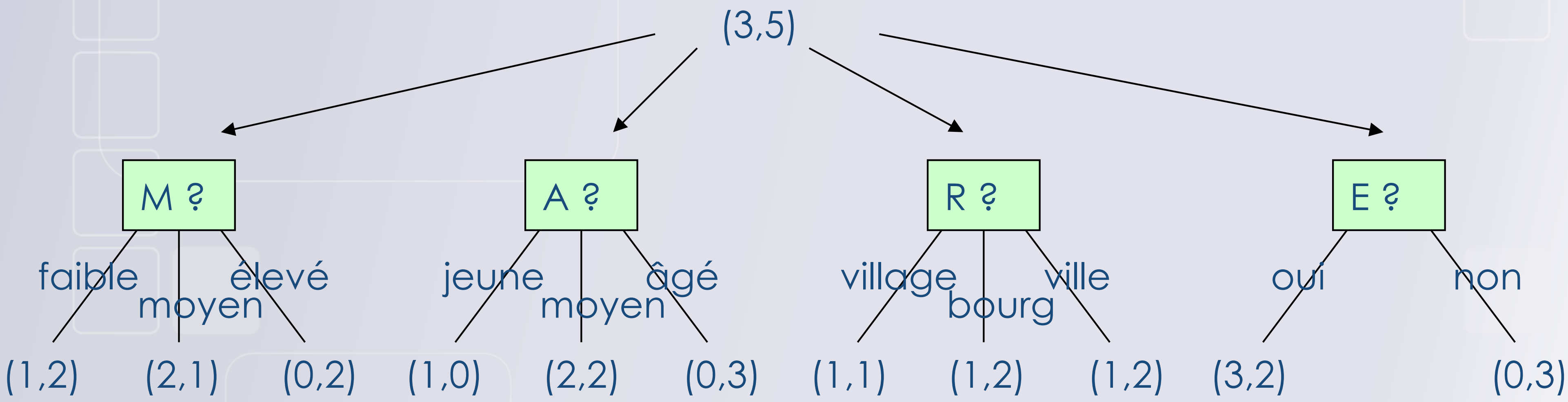
	M	A	R	E	I
1	moyen	moyen	village	oui	oui
2	élevé	moyen	bourg	non	non
3	faible	âgé	bourg	non	non
4	faible	moyen	bourg	oui	oui
5	moyen	jeune	ville	oui	oui
6	élevé	âgé	ville	oui	non
7	moyen	âgé	ville	oui	non
8	faible	moyen	village	non	non

Arbre de décision : Exemple

- **Racine de l'arbre (pas de test)**

  Etiquette (3,5) correspondant à : **3** oui et **5** non

- **Quel est premier test à réaliser ?**



- **Intuitivement :**
  - Le test sur R n'est pas discriminatoire
  - Le test sur A est interessant sur les branches jeune et âgé

Les arbres de décision

- **Quelles fonctions permettraient de réprésenter ces intuitions ?**

- **Fonctions qui seraient :**
  - Minimum lorsque le nœud est pur (tous les exemples sont dans une même classe
  - et Maximum lorsque les exemples sont équirépartis.

- **Exemples de fonctions possédant ces propriétés :**

  - **Entropie :** $$\text{Entropie}(p) = -\sum_{k=1}^{c} P(k/p) \times \log(P(k/p))$$

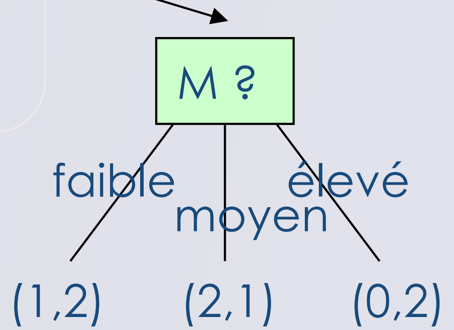
    Mesure le désordre en *thermodynamique*

  - **Fonction de Gini :** $$Gini(p) = 1 - \sum_{k=1}^{c} P(k/p)^2$$

Les arbres de décision

- **Exemple traité avec l'entropie :**

(3,5) $$Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0{,}954$$



$$Entropie(1) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0.918$$

$$Entropie(2) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \approx 0.918$$

$$Entropie(3) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

Les arbres de décision

- **Quelle fonction permettrait choisir un test ?**

  - **Fonction de gain :** $$Gain(p,test) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)$$

    p : position

    P_j: proportion d'éléments de S à la position p qui vont en position p_j

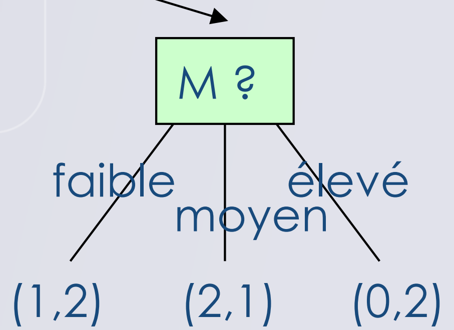
    $i(p)$:Entropie(p) ou Gini(p)

  - **Le test choisi est celui qui possède le gain le plus grand**

Les arbres de décision

- **Exemple traité avec l'entropie :**

(3,5) $Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0,954$



$$Entropie(1) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0.918$$

$$Entropie(2) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \approx 0.918$$

$$Entropie(3) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

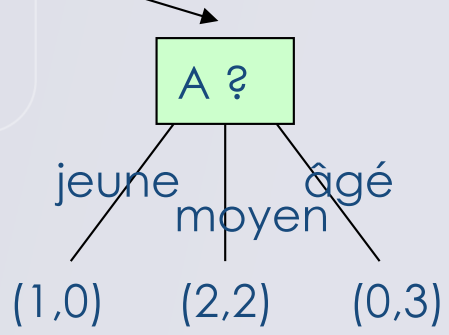
$$Gain(\in, M) = Entropie(\in) - \left(\frac{3}{8}Entropie(1) + \frac{3}{8}Entropie(2) + \frac{2}{8}Entropie(3)\right)$$

$$= Entropie(\in) - 0,688$$

Les arbres de décision

- **Exemple traité avec l'entropie :**

(3,5) $$Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0{,}954$$



$$Entropie(1) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right) = 0$$

$$Entropie(2) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$

$$Entropie(3) = -\frac{0}{3}\log\left(\frac{0}{3}\right) - \frac{3}{3}\log\left(\frac{3}{3}\right) = 0$$

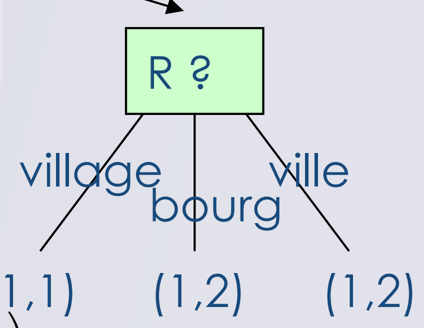
$$Gain(\in, A) = Entropie(\in) - \left(\frac{1}{8}Entropie(1) + \frac{4}{8}Entropie(2) + \frac{3}{8}Entropie(3)\right)$$

$$= Entropie(\in) - 0{,}5$$

Les arbres de décision

- **Exemple traité avec l'entropie :**

(3,5) $$Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0,954$$



$$Entropie(1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(2) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0,918$$

$$Entropie(3) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0,918$$

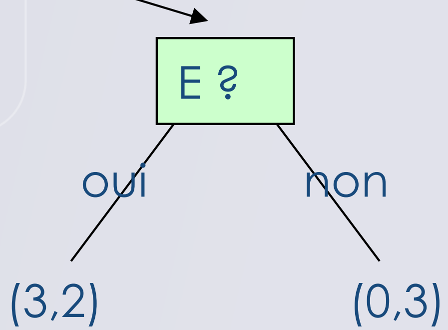
$$Gain(\in, R) = Entropie(\in) - \left(\frac{2}{8}Entropie(1) + \frac{3}{8}Entropie(2) + \frac{3}{8}Entropie(3)\right)$$

$$= Entropie(\in) - 0,938$$

Les arbres de décision

- **Exemple traité avec l'entropie :**

(3,5) $$Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0,954$$



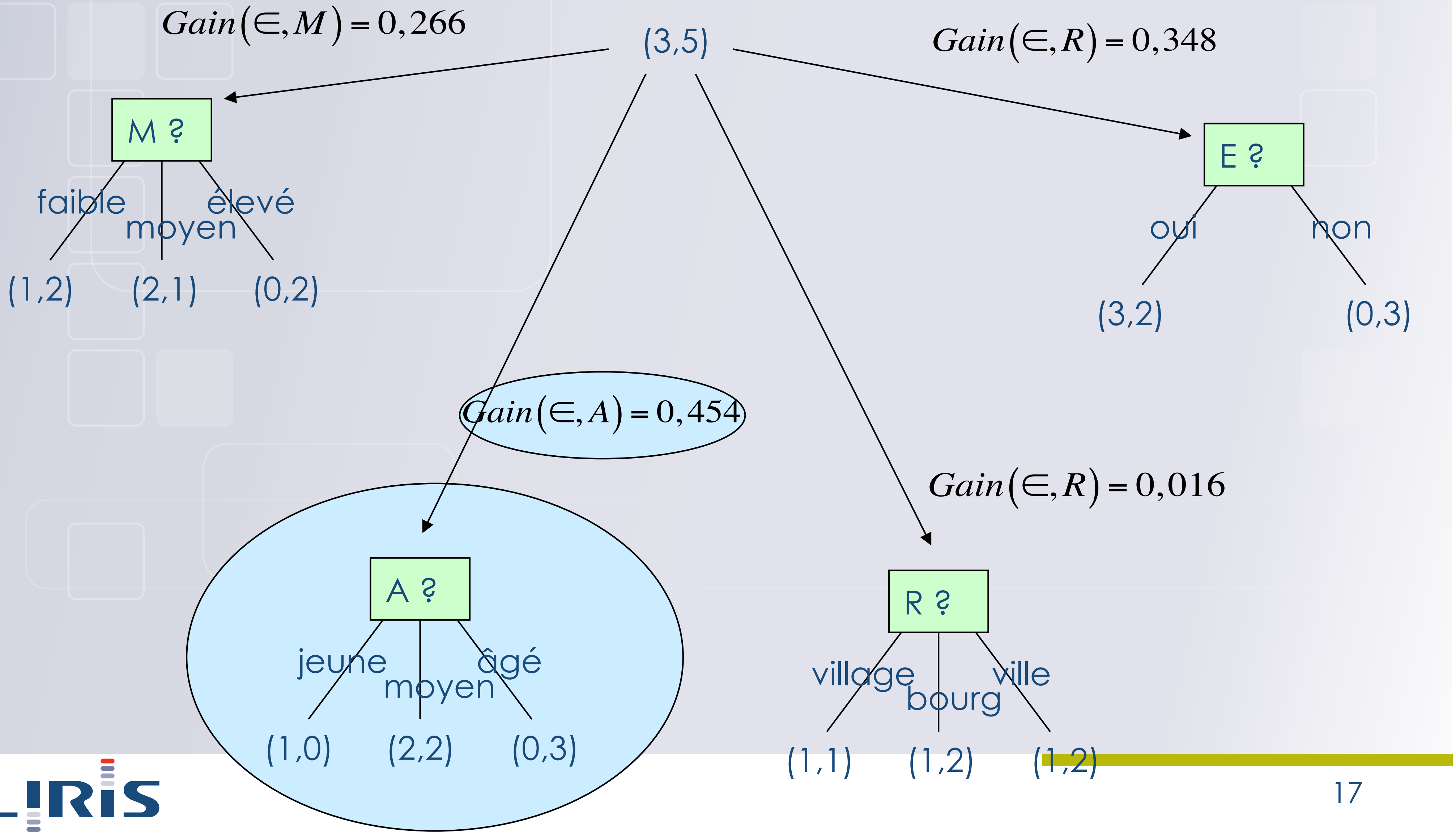
$$Entropie(1) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) \approx 0.970$$

$$Entropie(2) = -\frac{0}{3}\log\left(\frac{0}{3}\right) - \frac{3}{3}\log\left(\frac{3}{3}\right) = 0$$

$$Gain(\in, E) = Entropie(\in) - \left(\frac{5}{8}Entropie(1) + \frac{3}{8}Entropie(2)\right) = Entropie(\in) - 0,606$$

Les arbres de décision

- **Exemple traité avec l'entropie :**



Les arbres de décision

Algorithme générique de construction d'un arbre

entrée : échantillon S

début

Initialiser à l'arbre vide; la racine est le nœud courant

répéter

Décider si le nœud courant est terminal

si le nœud est terminal **alors**

Affecter une classe

sinon

Sélectionner un test et créer le sous-arbre

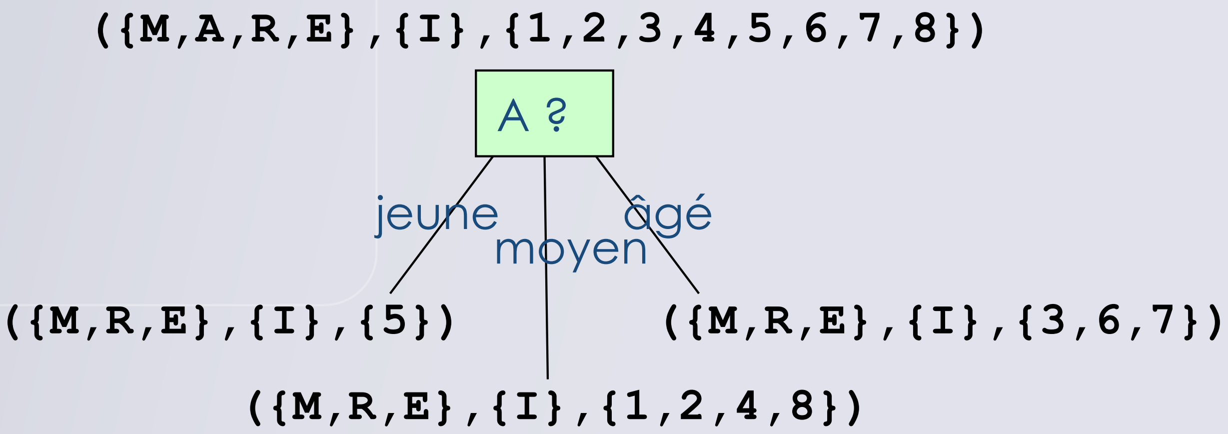
finsi

Passer au nœud suivant non exploré si il en existe

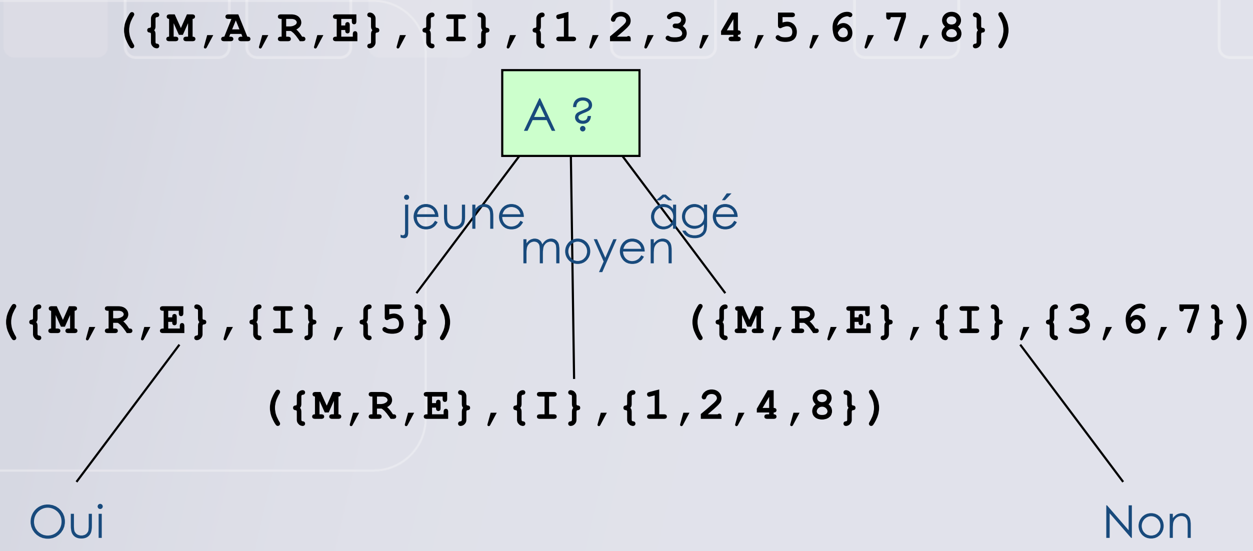
jusqu'à obtenir un arbre de décision (plus de nœud sans classe)

fin

Les arbres de décision



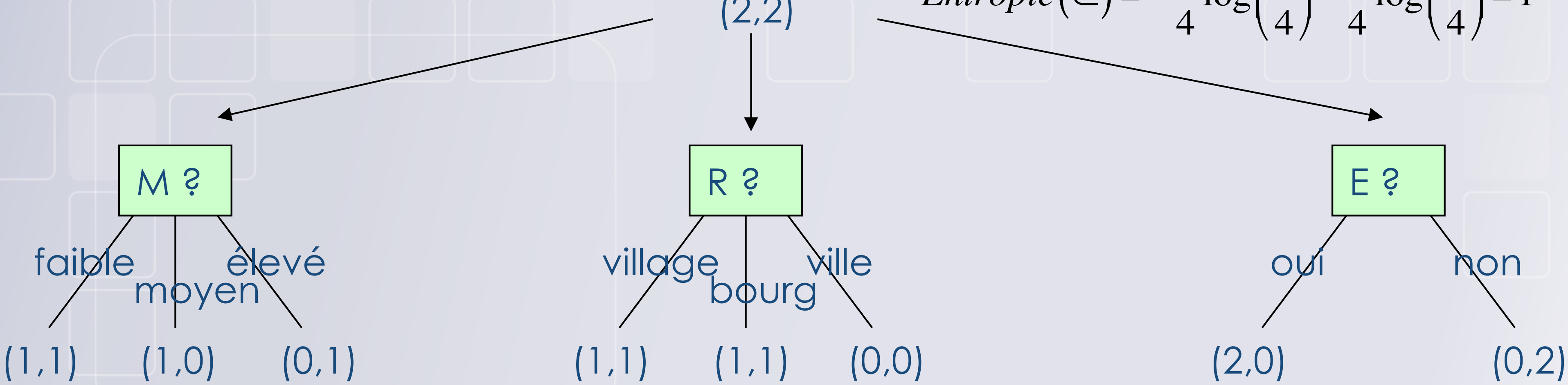
Les arbres de décision



	M	R	E	I
1	moyen	village	oui	oui
2	élevé	bourg	non	non
4	faible	bourg	oui	oui
8	faible	village	non	non

Les arbres de décision

$$Entropie(\in) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$



$$Entropie(1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(2) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right) = 0$$

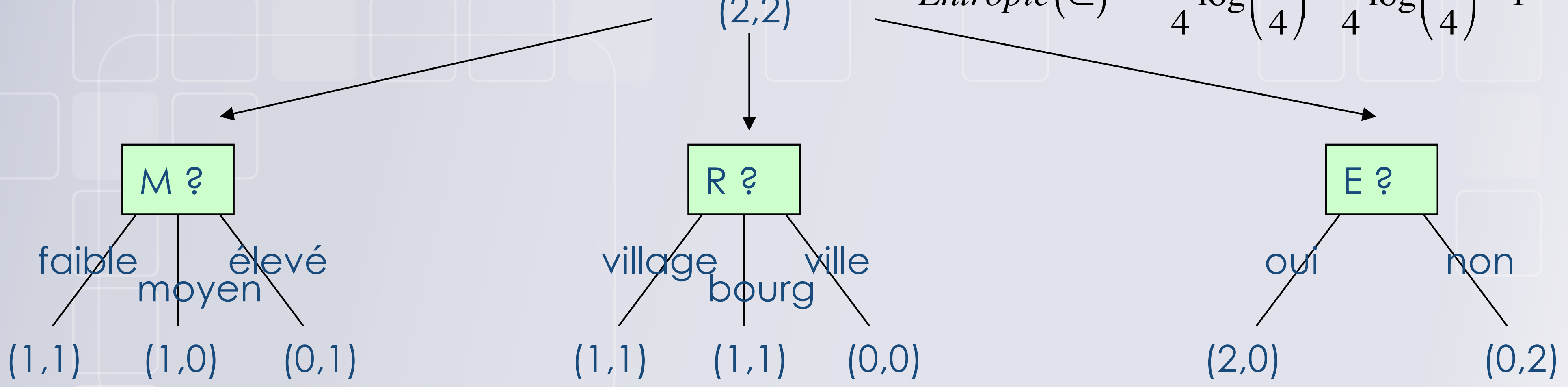
$$Entropie(3) = -\frac{0}{1}\log\left(\frac{0}{1}\right) - \frac{1}{1}\log\left(\frac{1}{1}\right) = 0$$

$$Gain(\in, M) = Entropie(\in) - \left(\frac{2}{4}Entropie(1) + \frac{1}{1}Entropie(2) + \frac{1}{1}Entropie(3)\right)$$

$$= Entropie(\in) - 0,5$$

Les arbres de décision

$$Entropie(\in) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$



$$Entropie(1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(2) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

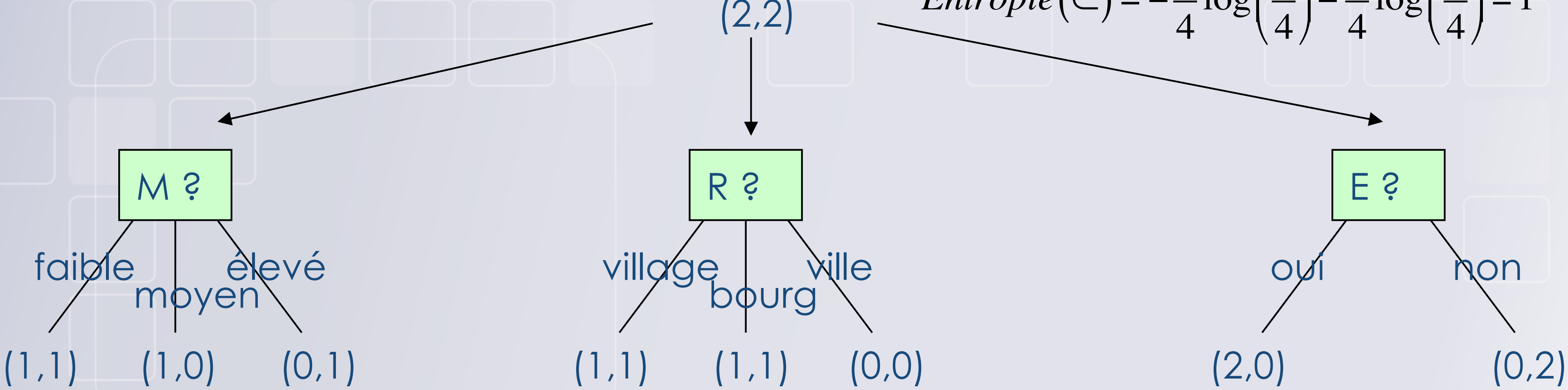
$$Entropie(3) = -\frac{0}{0}\log\left(\frac{0}{0}\right) - \frac{0}{0}\log\left(\frac{0}{0}\right) = 0$$

$$Gain(\in, R) = Entropie(\in) - \left(\frac{2}{4}Entropie(1) + \frac{2}{4}Entropie(2) + \frac{0}{4}Entropie(3)\right)$$

$$= Entropie(\in) - 1$$

Les arbres de décision

$$Entropie(\in) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$



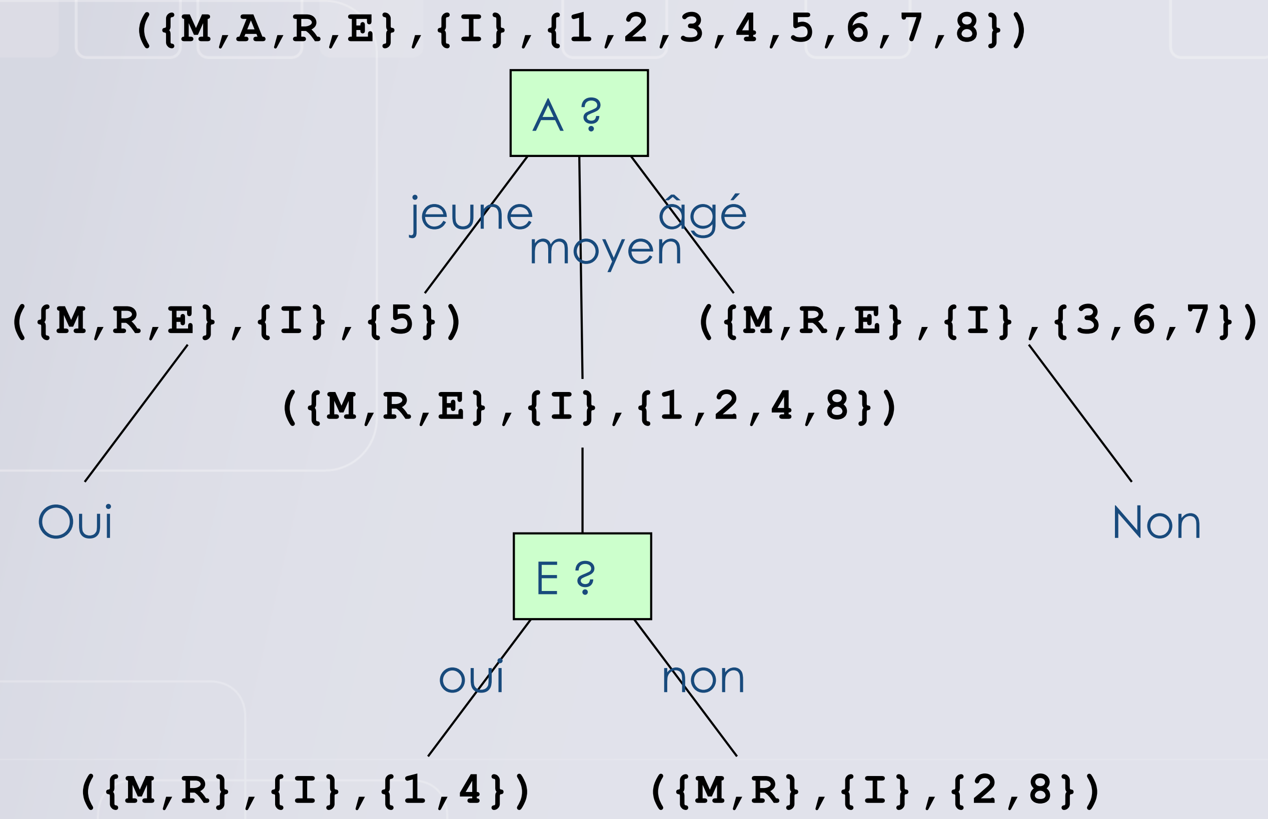
$$Entropie(1) = -\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log\left(\frac{0}{2}\right) = 0$$

$$Entropie(2) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

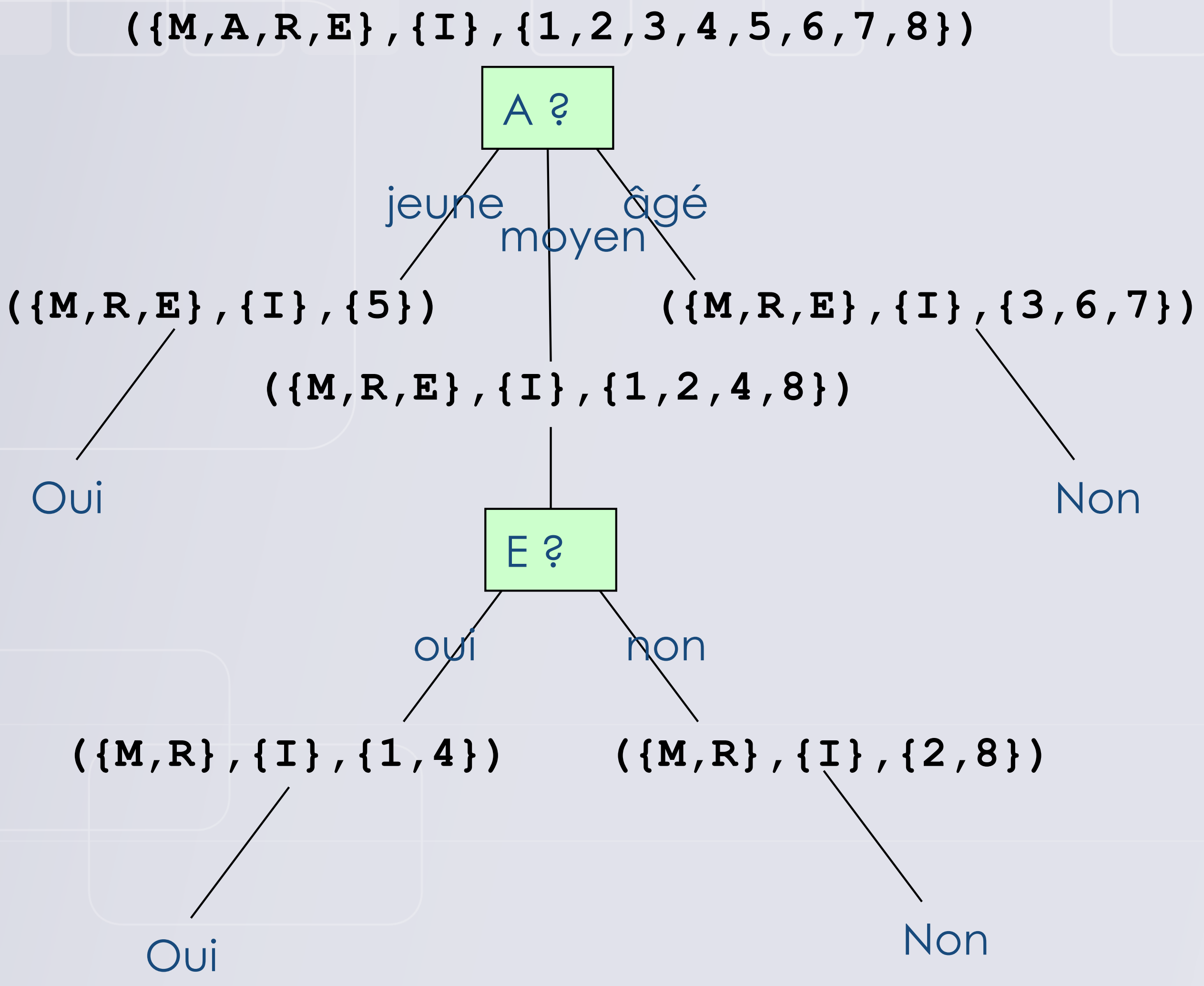
$$Gain(\in, E) = Entropie(\in) - \left(\frac{2}{4}Entropie(1) + \frac{2}{4}Entropie(2)\right)$$

$$= Entropie(\in) - 0$$

Les arbres de décision

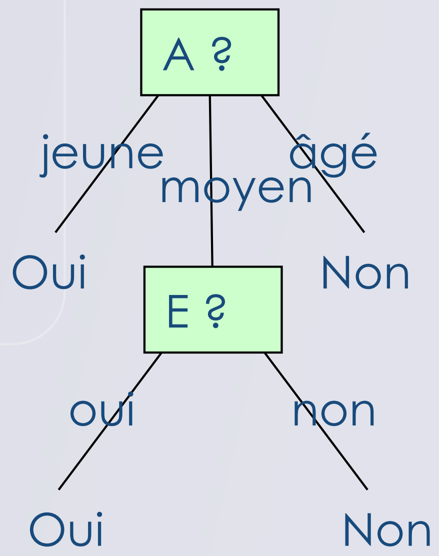


Les arbres de décision



Les arbres de décision

Arbre final



Évaluation d'un Arbre de Décision

	M	A	R	E	Re	Pr
1	moyen	moyen	village	oui	oui	oui
2	élevé	moyen	bourg	non	non	non
3	faible	âgé	bourg	non	non	non
4	faible	moyen	bourg	oui	oui	oui
5	moyen	jeune	ville	oui	oui	oui
6	élevé	âgé	ville	oui	non	non
7	moyen	âgé	ville	oui	non	non
8	faible	moyen	village	non	non	non
9	moyen	âgé	village	oui	oui	non
10	élevé	jeune	ville	non	oui	oui
11	faible	âgé	village	non	non	non
12	moyen	moyen	bourg	oui	oui	oui

		Pr	
		Non	Oui
Re	Non	1 (VN)	0 (FP)
	Oui	1 (FN)	2 (VP)

- **Accuracy =** 0.75
- **Erreur=** 0.25
- **Rappel (oui) = VP/(VP+FN)** 66%
- **Précision (oui) = VP/(VP+FP)** 100%
- **F-mesure (oui) =** 79.5%