

TD 4 en Fouille de données

On veut prédire la récidive (refaire une infraction) de détenus d'une prison selon leurs caractéristiques. Les détenus sont décrits par leur genre (H/F), leur catégorie d'âge (C1, C2, C3), le nombre d'incarcérations, et s'ils travaillent. Le tableau suivant décrit les caractéristiques de 10 détenus

Genre	Age	Nombre d'incarcération	Travail	Récidive
H	C1	1	Oui	Oui
H	C2	1	Oui	Non
H	C1	2	Non	Oui
H	C3	1	Oui	Non
H	C1	3+	Non	Oui
F	C1	1	Non	Non
F	C2	1	Oui	Non
F	C3	3+	Non	Oui
F	C3	2	Oui	Non
F	C3	1	Non	Non

Travail demandé

1. En utilisant l'algorithme ID3 et l'entropie comme critère de sélection des attributs, donner **la racine** de l'arbre de décision et les branches en attente et les détails des calculs.
2. On veut arrêter le développement de l'arbre par pré-élagage. Donner l'arbre complet.
3. Quelle est la qualité de cet arbre en termes de précision et de rappel?
4. Est-il possible d'améliorer la précision de cet arbre? Justifier la réponse.

Annexe: la première ligne donne les paires de probabilités (ex: 1/4 et 3/4) et la deuxième l'entropie correspondant.

Probabilités. (%)	1/4 et 3/4	1/5 et 4/5	2/5 et 3/5	1/6 et 5/6	2/6 et 4/6	1/7 et 6/7	2/7 et 5/7	3/7 et 4/7
Entropie	0.81	0.72	0.97	0.66	0.91	0.58	0.87	0.99