

TD Arbre de Décision
ESP-UCAD

Dr. Mamadou Camara
mamadou.camara@esp.sn

2022-2023

Table des matières

1	Apprentissage supervisé :	
	Les arbres de décision	2
1.1	TD rpart	2
1.2	TD Gain	4
1.3	Mode de transport	4
2	Annexe	5
2.1	Mode de transport [Master Recherche SI]	5
2.2	TD sur le critère de gain [DIC]	6
2.3	DS DM 2021 Semestre 2 : Partitionnements (15 points)	7

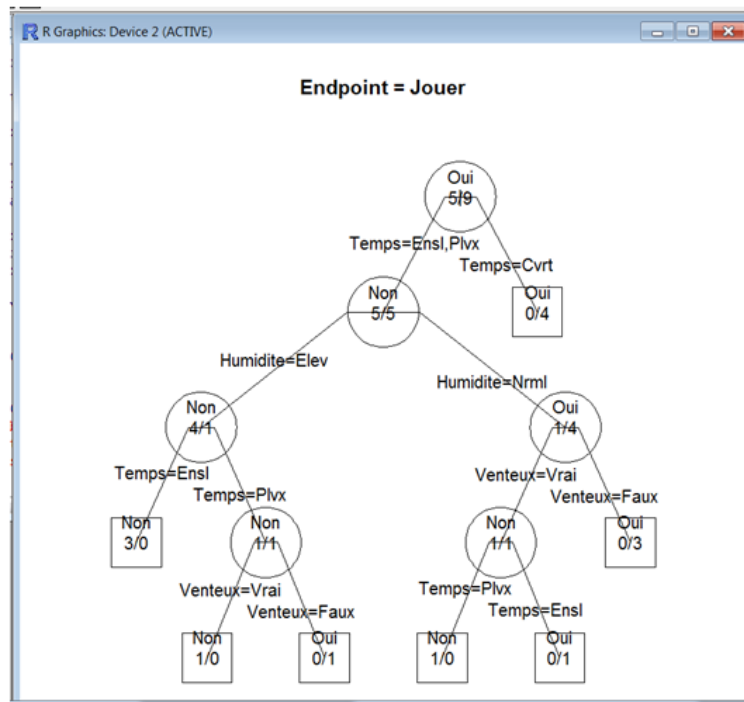
Chapitre 1

Apprentissage supervisé : Les arbres de décision

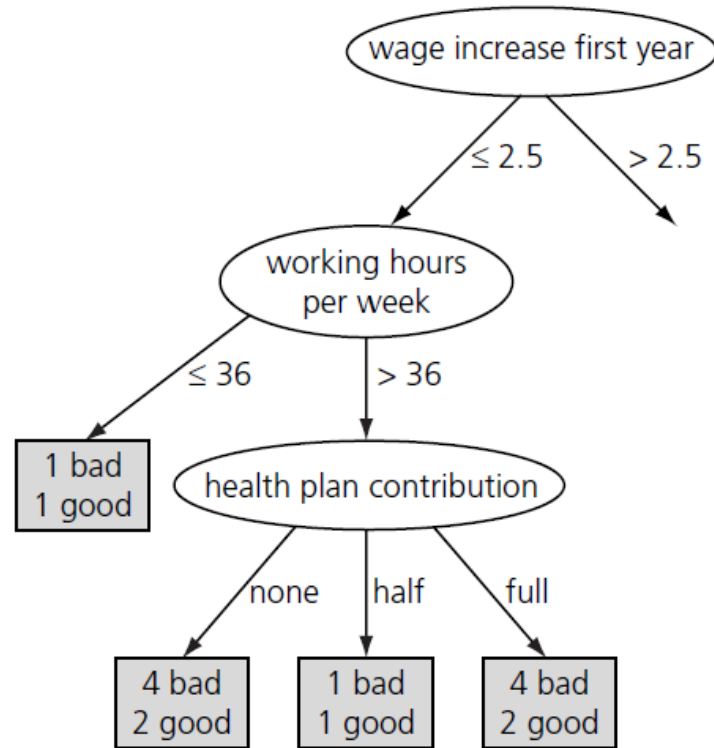
1.1 TD rpart

Le package rpart de R permet de construire l'arbre ci-dessous.

— Définir et interpréter les informations contenues dans le sous-arbre nsplit=2 de cet arbre .



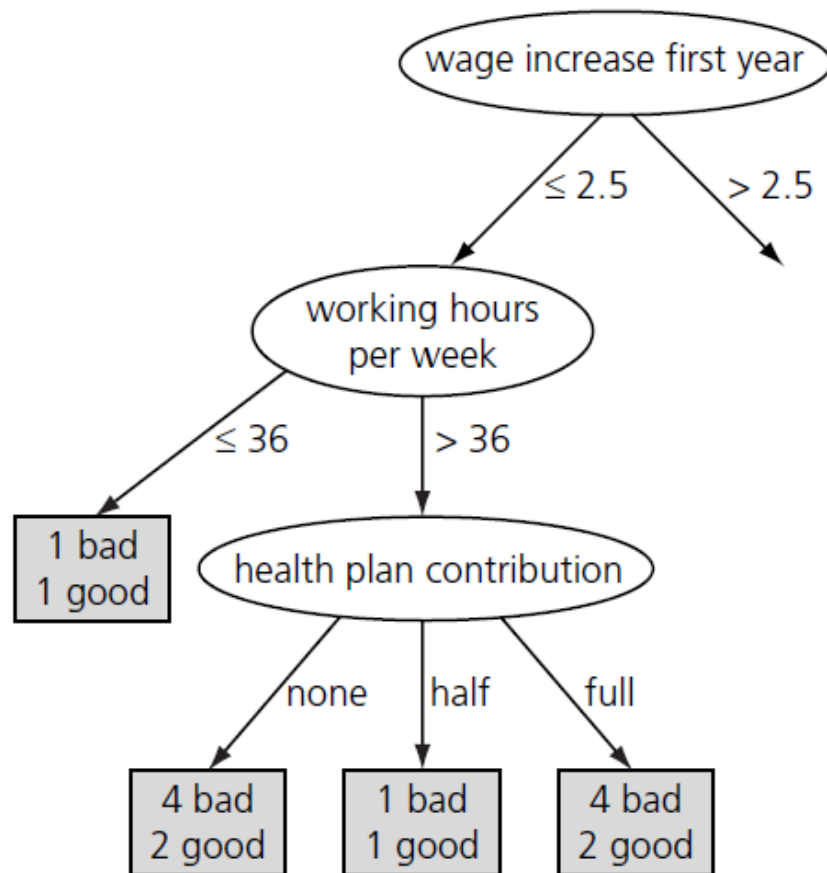
— Définir et interpréter les informations contenues dans l'arbre suivant.



— Expliquer les différences de représentation entre ces deux arbres de décisions.

1.2 TD Gain

Calculer le gain d'information au niveau du partitionnement qui utilise la variable "Working hours per week" (i.e. nombre d'heures travaillées par semaine).



1.3 Mode de transport

— Calculer un gain d'information avec l'attribut "car ownership" avec l'hypothèse d'un arbre binaire.

Attributes				Classes
Gender	Car ownership	Travel Cost (\$)/km	Income Level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Chapitre 2

Annexe

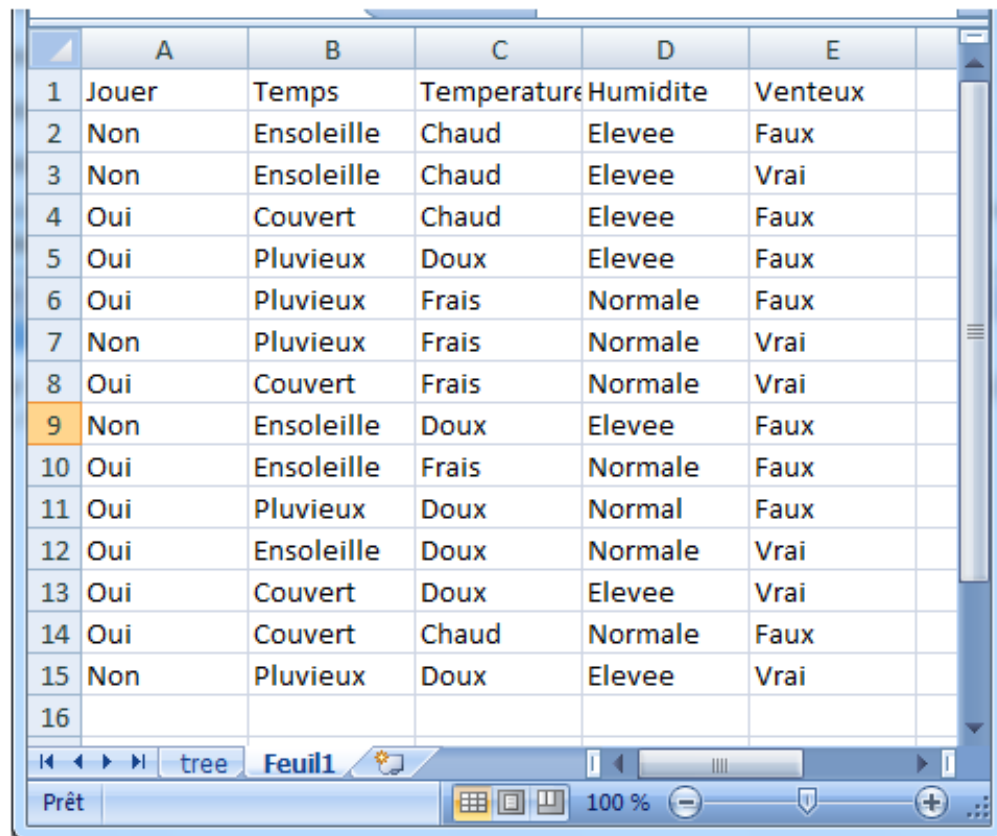
2.1 Mode de transport [Master Recherche SI]

— Calculer un gain d'information avec l'attribut "niveau de revenu" avec l'hypothèse d'un arbre binaire.

Attributes				Classes
Gender	Car ownership	Travel Cost (\$)/km	Income Level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

2.2 TD sur le critère de gain [DIC]

Considérons un arbre à construire à partir des données suivantes :



	A	B	C	D	E
1	Jouer	Temps	Temperature	Humidite	Venteux
2	Non	Ensoleille	Chaud	Elevee	Faux
3	Non	Ensoleille	Chaud	Elevee	Vrai
4	Oui	Couvert	Chaud	Elevee	Faux
5	Oui	Pluvieux	Doux	Elevee	Faux
6	Oui	Pluvieux	Frais	Normale	Faux
7	Non	Pluvieux	Frais	Normale	Vrai
8	Oui	Couvert	Frais	Normale	Vrai
9	Non	Ensoleille	Doux	Elevee	Faux
10	Oui	Ensoleille	Frais	Normale	Faux
11	Oui	Pluvieux	Doux	Normal	Faux
12	Oui	Ensoleille	Doux	Normale	Vrai
13	Oui	Couvert	Doux	Elevee	Vrai
14	Oui	Couvert	Chaud	Normale	Faux
15	Non	Pluvieux	Doux	Elevee	Vrai
16					

En vous positionnant à la racine de l'arbre, calculer le gain d'information pour la variable Humidité. L'arbre n'est pas binaire.

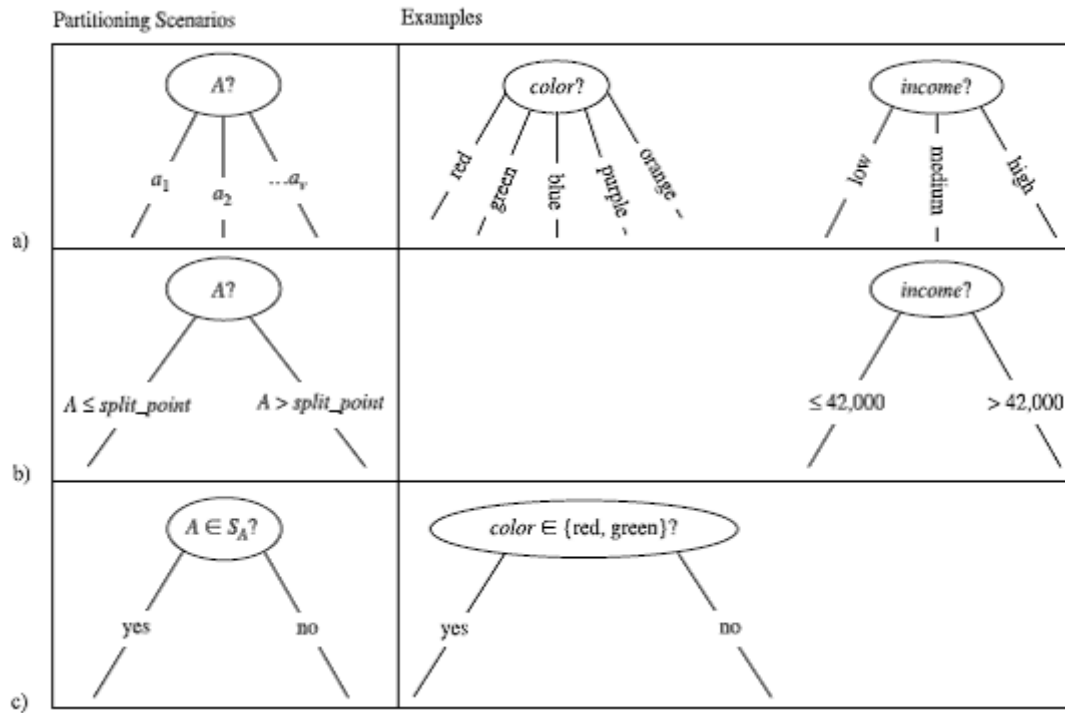
1. Dessiner le partitionnement avec la variable Humidité
2. Calculer l'entropie sur la racine
3. Calculer l'entropie sur les sous-nœuds
4. Calculer le gain

2.3 DS DM 2021 Semestre 2 : Partitionnements (15 points)

Considérons les quatre scénarios de partitionnement ci-après :

1. la variable est discrète & l'arbre est binaire
2. la variable est discrète & l'arbre n'est pas binaire
3. la variable est continue & faire la discrétisation
4. la variable est continue & ne pas faire la discrétisation

Considérons les trois lignes de la figure ci-après (i.e. a, b & c).



Traiter les questions suivantes :

1. Définir chacun des éléments qui se trouvent sur la figure :
 - A, color, a1, ar, red, ..., orange, income, low, medium, high
 - A, split_point, income, 42000
 - A, SA, color, red, green
2. Faire la correspondance entre les scénarios listés ci-dessus et les lignes présentes dans la figure.
3. Considérons maintenant la règle générale selon laquelle une variable explicative correspond à une option de partitionnement. Dire, pour chaque scénario, si c'est la règle générale qui s'applique ou une exception à cette règle.
4. Éventuellement, expliquer exception en question en utilisant un exemple issu ou non de l'énoncé de cet exercice.