DATA MINING

CLASSIFICATION SUPERVISÉE

MODÈLES À BASE DE RÈGLES DÉCISIONNELLES :

ARBRES DE DÉCISION

Mohamed Heny SELMI- Wiem Trabelsi

EXEMPLE – INFORMATION QUALITATIVE

client	*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

Une banque dispose des informations suivantes sur un ensemble de clients:

M **: moyenne des montants sur le compte client.**

A **: tranche d'âge du client.**

R **: localité de résidence du client.**

E **: valeur oui si le client a un niveau d'études supérieures.**

I : classe oui correspond à un client qui effectue une consultation de ses comptes bancaires en utilisant Internet

Quelle est la variable à mettre comme racine de l'arbre?

A ?

E ?

M ?

R ?

PRINCIPE ALGORITHMIQUE DES ARBRES DE DÉCISIONS

Procédure *construire-arbre(X)*

SI tous les individus I appartiennent à la même modalité de la variable décisionnelle

ALORS créer un nœud feuille portant le nom de cette classe : Décision

SINON

- ✔ choisir le meilleur attribut pour créer un nœud // l'attribut qui sépare le mieux
- ✔ le test associé à ce nœud sépare X en des branches : X_d……X_g
  - ✔ *construire-arbre(X_d)*
  - …
  - …
  - …
  - ✔ *construire-arbre(X_g)*

FIN

CHOIX DU MEILLEUR ATTRIBUT POUR CRÉER UN NŒUD

- Il existe plusieurs méthodes pour choisir le meilleur attribut à placer dans un nœud :

✔ Algorithme C4.5, C5.0

✔ CHAID Chi-squared Automatic Interaction Detector

✔ ID3 entropie de Shannon

✔ **CART Classification and regression trees : Indice de GINI**

- **l'indice de GINI est le meilleur moyen pour la construction de l'arbre car il est le seul indice qui répond aux questions suivantes :**

✔ Comment choisir la variable à segmenter parmis les variables explicatives disponibles ?

✔ Lorsque la variable est continue, comment déterminer le seuil de coupe ?

✔ Comment déterminer la bonne taille de l'arbre ?

ALGORITHME DE CART

- ✔ Parmi les plus performants et plus répandus
- ✔ Accepte tout type de variables
- ✔ Utilise le **Critère de séparation : Indice de Gini** $I = 1 - \sum_{i}^{n} f_i^2$

Avec n : nombre de classes à prédire

f_i : fréquence de la classe dans le nœud

- ✔ *Plus l'indice de Gini est bas, plus le nœud est pure*
- ✔ En séparant 1 nœud en 2 nœuds fils on cherche la plus grande hausse de la pureté
- ✔ La variable la plus discriminante doit maximiser

IG(avant séparation)-[IG($fils_1$)+…. …. +IG($fils_n$)]

CALCUL DE L'INDICE DE GINI

Indice de Gini avant séparation au NIVEAU DE LA RACINE :

8 clients

- **I=oui : 3 clients**
- **I=non : 5 clients**

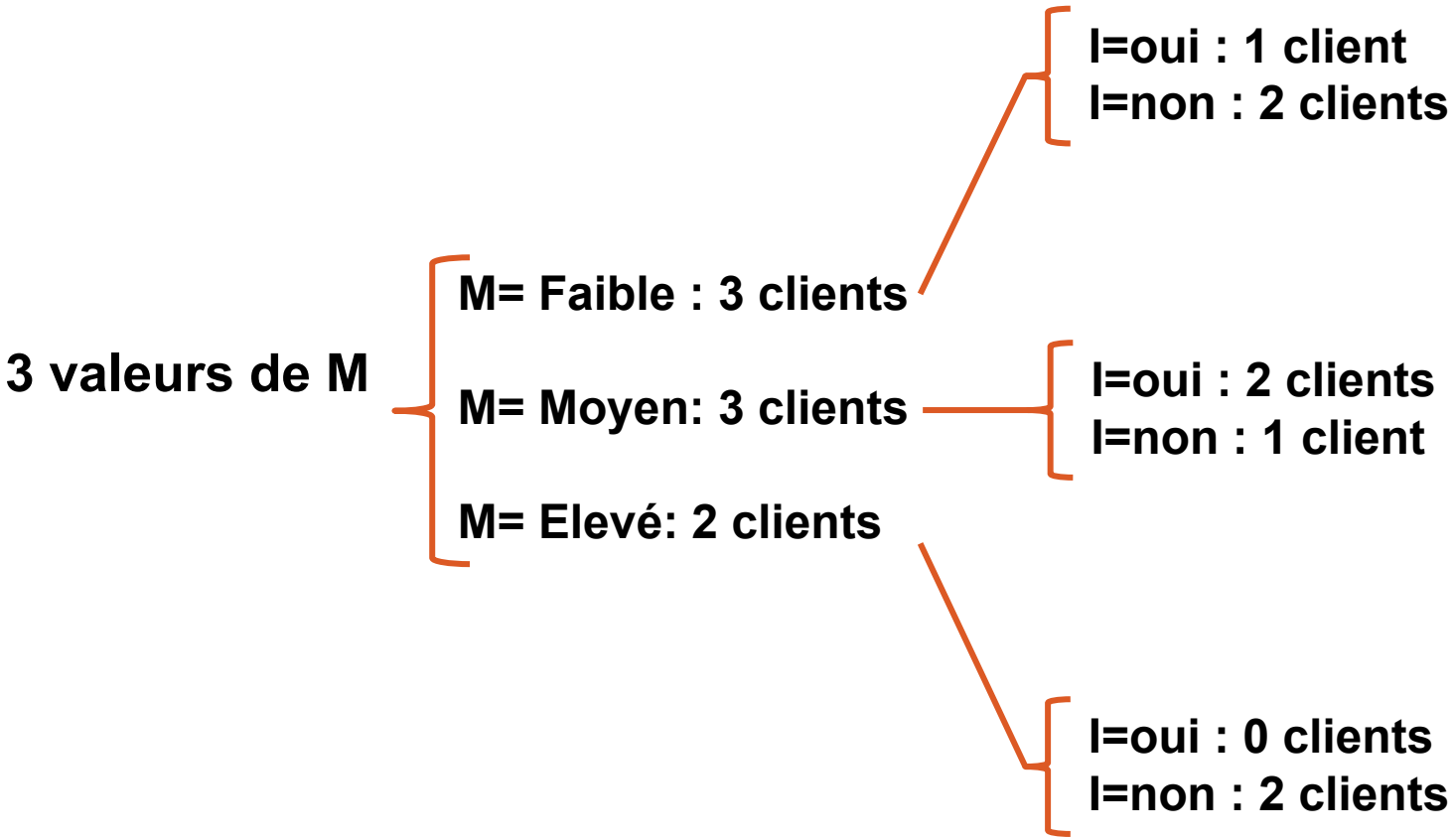
$$IG(\text{avant séparation}) = 1 - \left((3/8)^2 + (5/8)^2 \right) = 0.46875$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable M (Moyenne des montants sur le compte client):



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils M = Faible :

3 clients

- **I=oui : 1 client**
- **I=non : 2 clients**

$$IG(M=Faible) = 1 - ((1/3)^2 + (2/3)^2) = 0.4444444$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils M = Moyen :

3 clients { **I=oui : 2 clients**, **I=non : 1 client** }

$$IG(M=Moyen) = 1 - ((2/3)^2 + (1/3)^2) = 0.4444444$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils M = Elevé :

2 clients

- **I=oui : 0 clients**
- **I=non : 2 clients**

$$IG(M=Elevé) = 1 - ((0/2)^2 + (2/2)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de M:

$$IG(\text{avant séparation})-[IG(M=\text{Faible})+IG(M=\text{Moyen})+IG(M=\text{Elevé})]$$

$$=$$

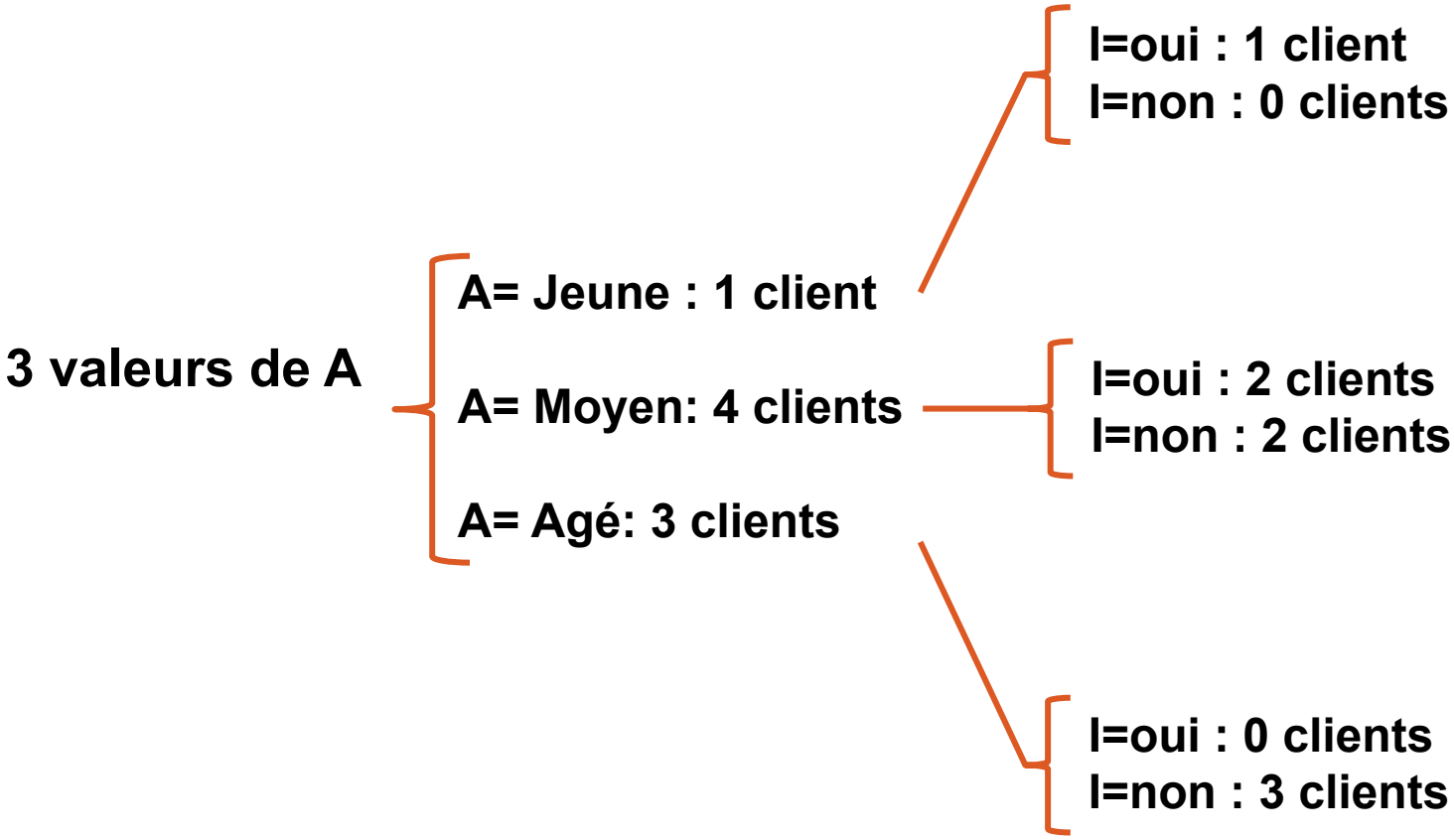
$$0.46875 - [0.4444444 + 0.4444444 + 0]$$

$$=$$

$$-0.4201388$$

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable A (Tranche d'âge du client):



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils A = Jeune :

1 client

- **I=oui : 1 client**
- **I=non : 0 clients**

$$IG(A=Jeune) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils A = Moyen :

4 clients { **I=oui : 2 clients**, **I=non : 2 clients** }

$$IG(A=Moyen) = 1 - ((2/4)^2 + (2/4)^2) = 0.5$$

Fréquence des I = oui — **Fréquence des I = non**

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils A = Agé :

3 clients

- **I=oui : 0 clients**
- **I=non : 3 clients**

$$IG(A=Agé) = 1 - ((0/3)^2 + (3/3)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de A:

$$IG(\text{avant séparation}) - [IG(A=\text{Jeune}) + IG(A=\text{Moyen}) + IG(A=\text{Agé})]$$

$$=$$

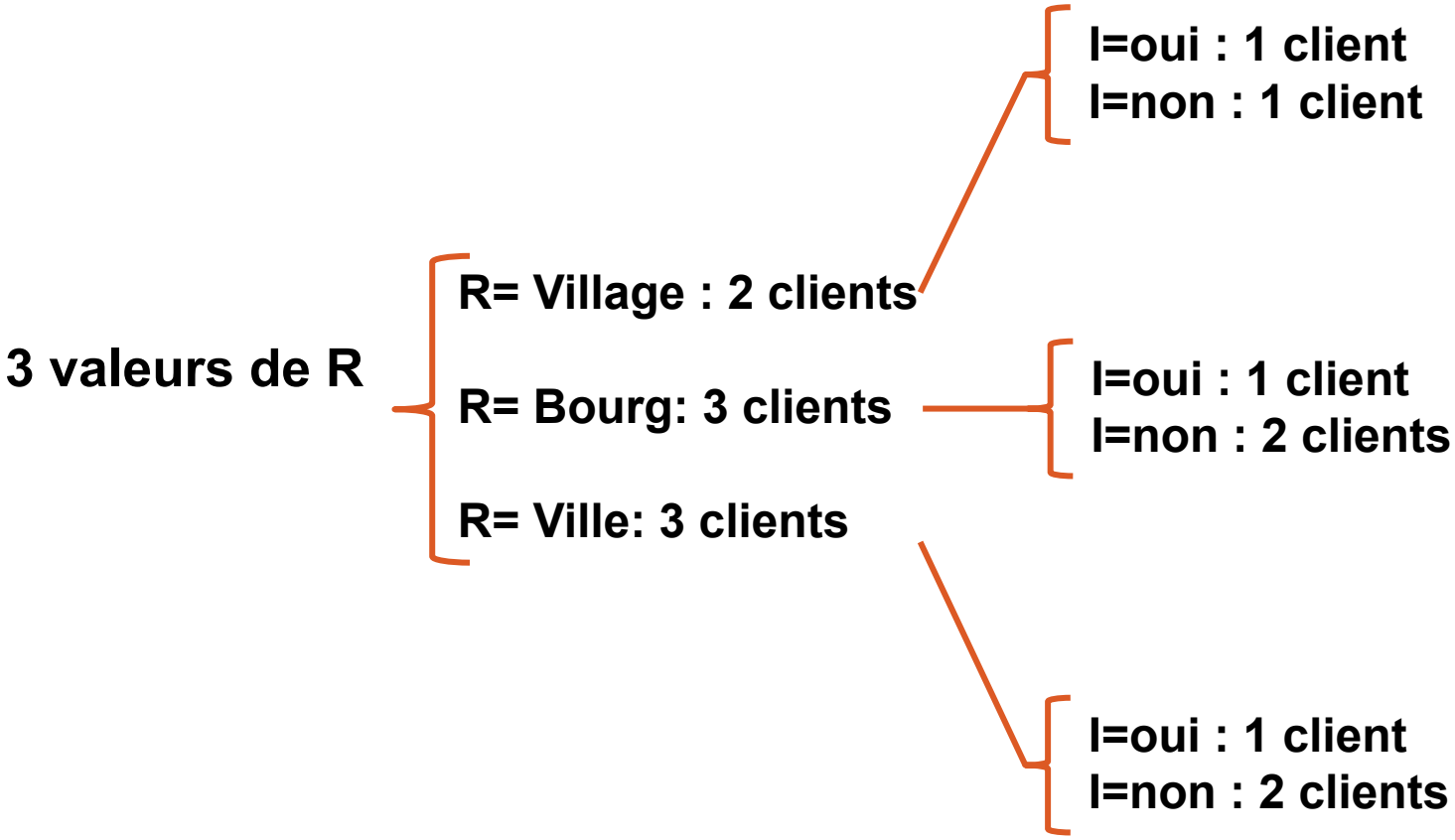
$$0.46875 - [0 + 0.5 + 0]$$

$$=$$

$$-0.03125$$

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable R(Localité de résidence du client):



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **R= Village** :

2 clients

- I=oui : 1 client
- I=non : 1 client

$$IG(R= Village) = 1 - ((1/2)^2 + (1/2)^2) = 0.5$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **R= Bourg** :

3 clients

- I=oui : 1 client
- I=non : 2 clients

$$IG(R= Bourg) = 1 - ((1/3)^2 + (2/3)^2) = 0.4444444$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils R= Ville:

3 clients

- **I=oui : 1 client**
- **I=non : 2 clients**

$$IG(R=Ville) = 1 - ((1/3)^2 + (2/3)^2) = 0.4444444$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de R:

$$IG(\text{avant séparation}) - [IG(R=\text{Village}) + IG(R=\text{Bourg}) + IG(R=\text{Ville})]$$

$$=$$

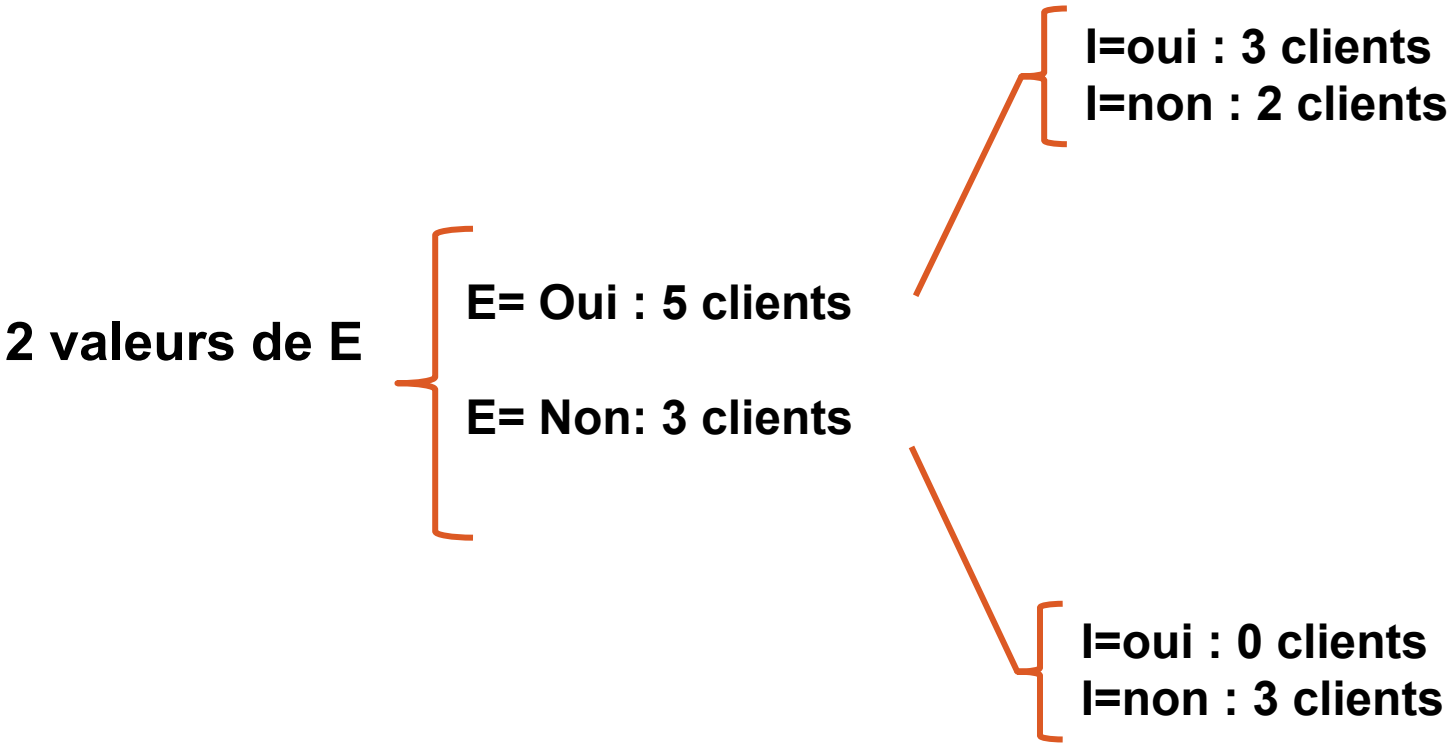
$$0.46875 - [0.4444444 + 0.5 + 0.4444444]$$

$$=$$

$$-0.9201388$$

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable E(Niveau d'études du client):



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils E= Oui :

5 clients

- **I=oui : 3 clients**
- **I=non : 2 clients**

$$IG(E=Oui) = 1 - ((3/5)^2 + (2/5)^2) = 0.48$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils E= Non :

3 clients

- **I=oui : 0 clients**
- **I=non : 3 clients**

$$IG(E=Non) = 1 - ((0/3)^2 + (3/3)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de E:

$$IG(\text{avant séparation}) - [IG(E=Oui) + IG(E=Non)]$$

$$=$$

$$0.46875 - [0.48+0]$$

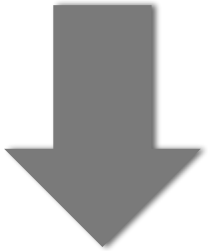
$$=$$

$$-0.01125388$$

PREMIER RESULTAT DE L'INDICE DE GINI

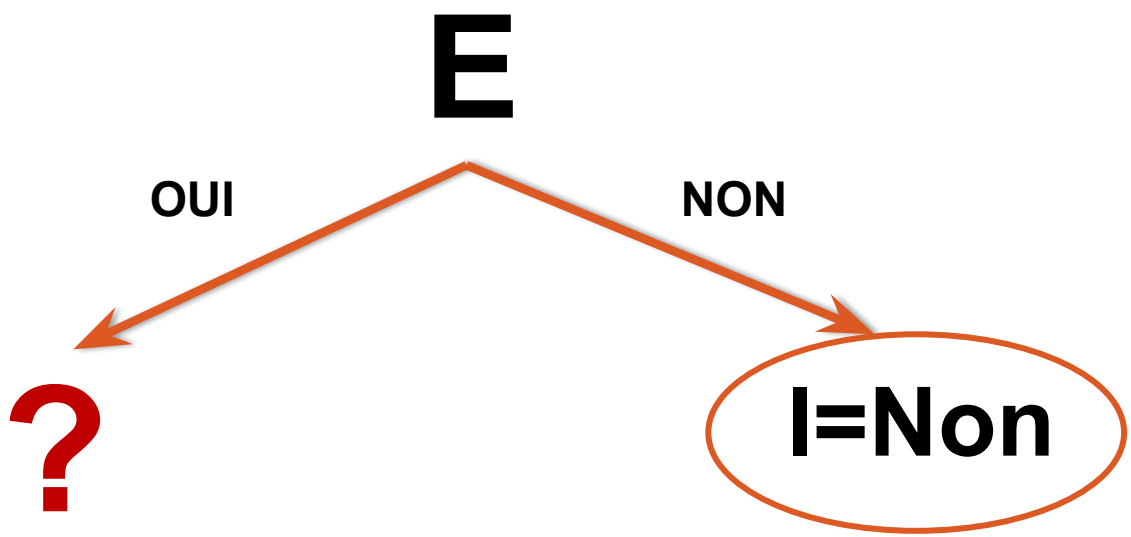
La variable la plus séparatrice est celle qui maximise :

$$IG(\text{avant séparation})-[IG(fils_1)+IG(fils_2)+\ldots\ldots+IG(fils_n)]$$



E

CONSTRUCTION DE L'ARBRE



CALCUL DE L'INDICE DE GINI : E=OUI

Indice de Gini avant séparation avec E = Oui :

5 clients

- **I=oui : 3 clients**
- **I=non : 2 clients**

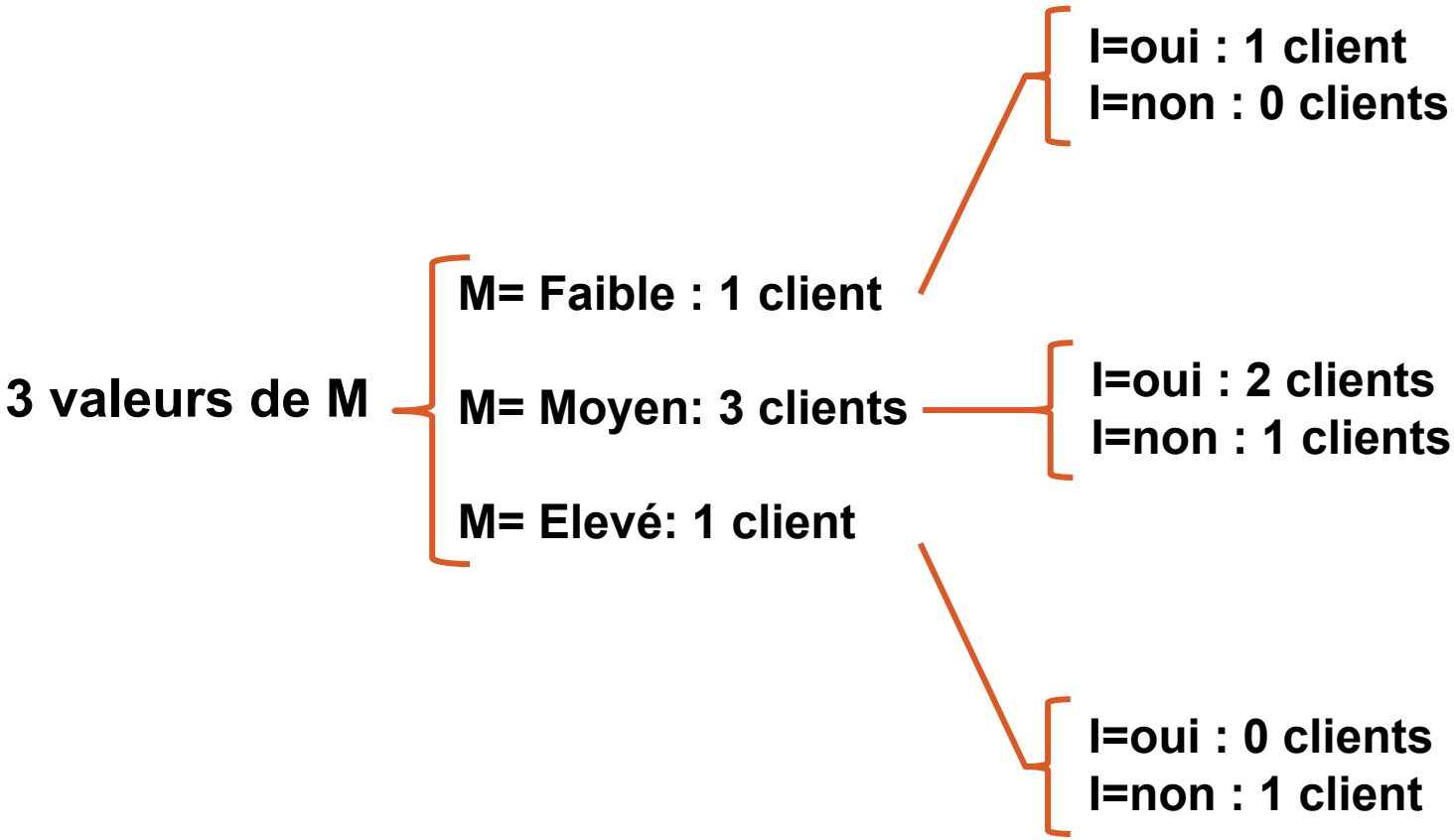
$$IG(\text{avant séparation}_1) = 1 - \left((3/5)^2 + (2/5)^2 \right) = 0.48$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable M (Moyenne des montants sur le compte client) avec E=Oui:



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **M = Faible & E = Oui** :

1 client

- I=oui : 1 client
- I=non : 0 clients

$$IG(M=Faible \ \& \ E=Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **M = Moyen & E = Oui** :

3 clients

- I=oui : 2 clients
- I=non : 1 client

$$IG(M=Moyen \ \& \ E=Oui) = 1 - ((2/3)^2 + (1/3)^2) = 0.4444444$$

- $(2/3)^2$: Fréquence des I = oui
- $(1/3)^2$: Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils M = Elevé & E = Oui:

1 client { **I=oui : 0 clients**
I=non : 1 client

$$IG(M=Elevé \ \& \ E=Oui) = 1 - ((0/1)^2 + (1/1)^2) = 0$$

Fréquence des I = oui — **Fréquence des I = non**

CALCUL DE L'INDICE DE GINI

Indice de Gini de M avec E=Oui :

$$IG(\text{avant séparation}_1)-[IG(M=\text{Faible})+IG(M=\text{Moyen})+IG(M=\text{Elevé})]$$

$$=$$

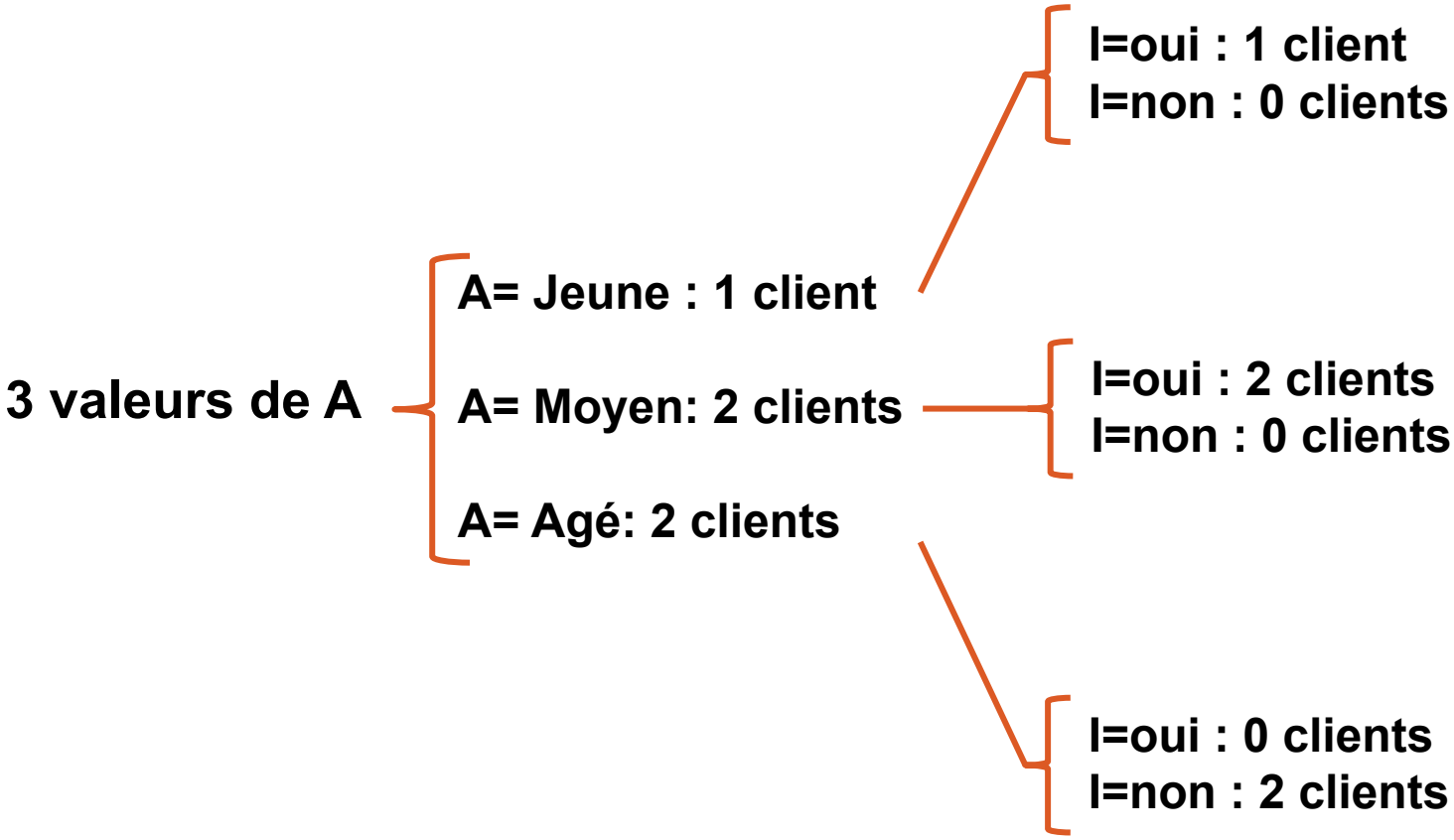
$$0.48 - [0 + 0.4444444 + 0]$$

$$=$$

$$0.0355556$$

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable A (Tranche d'âge du client) avec E=Oui :



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **A = Jeune & E = Oui** :

1 client

- I=oui : 1 client
- I=non : 0 clients

$$IG(A=Jeune \ \& \ E = Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **A = Moyen & E = Oui** :

2 clients

- I=oui : 2 clients
- I=non : 0 clients

$$IG(A=Moyen \ \& \ E = Oui) = 1 - ((2/2)^2 + (0/2)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **A = Agé & E = Oui** :

2 clients

- I=oui : 0 clients
- I=non : 2 clients

$$IG(A=Agé \ \& \ E = Oui) = 1 - ((0/2)^2 + (2/2)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de A avec E=Oui:

$$IG(\text{avant séparation}_1)-[IG(A=\text{Jeune})+IG(A=\text{Moyen})+IG(A=\text{Agé})]$$

$$=$$

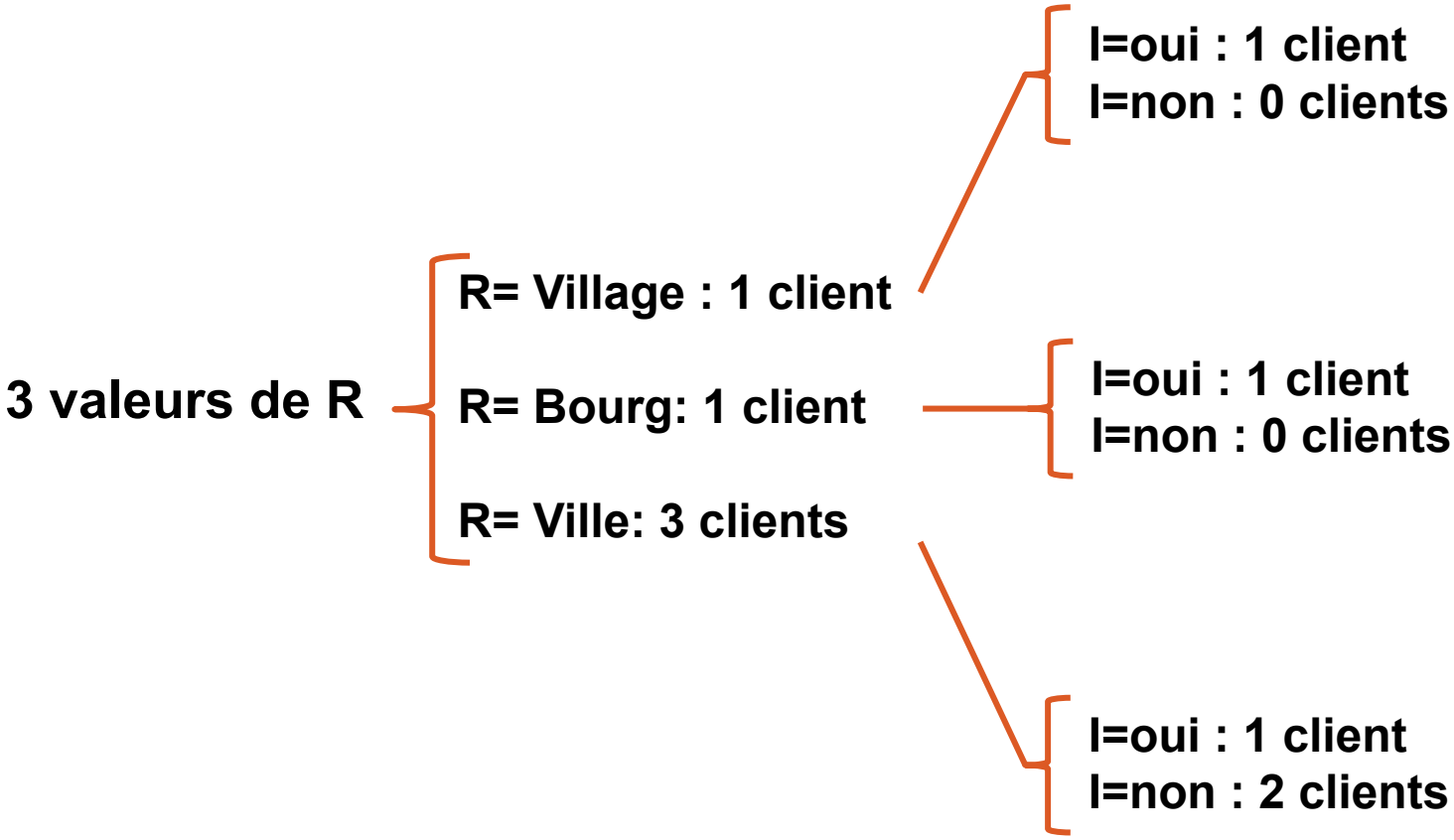
$$0.48 - [0 + 0 + 0]$$

$$=$$

$$\boxed{0.48}$$

CALCUL DE L'INDICE DE GINI

Indice de Gini de la variable R(Localité de résidence du client) avec E=Oui :



CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **R= Village & E = Oui** :

1 clients

- I=oui : 1 client
- I=non : 0 clients

$$IG(R= Village \ \& \ E = Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils R= Bourg & E = Oui :

1 client

- **I=oui : 1 client**
- **I=non : 0 clients**

$$IG(R= Bourg \ \& \ E = Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de fils **R= Ville & E = Oui** :

3 clients

- I=oui : 1 client
- I=non : 2 clients

$$IG(R=Ville\ \&\ E=Oui) = 1 - ((1/3)^2 + (2/3)^2) = 0.4444444$$

Fréquence des I = oui

Fréquence des I = non

CALCUL DE L'INDICE DE GINI

Indice de Gini de R avec E=Oui: :

$$IG(\text{avant séparation}_1)-[IG(R=Village)+IG(R=Bourg)+IG(R=Ville)]$$

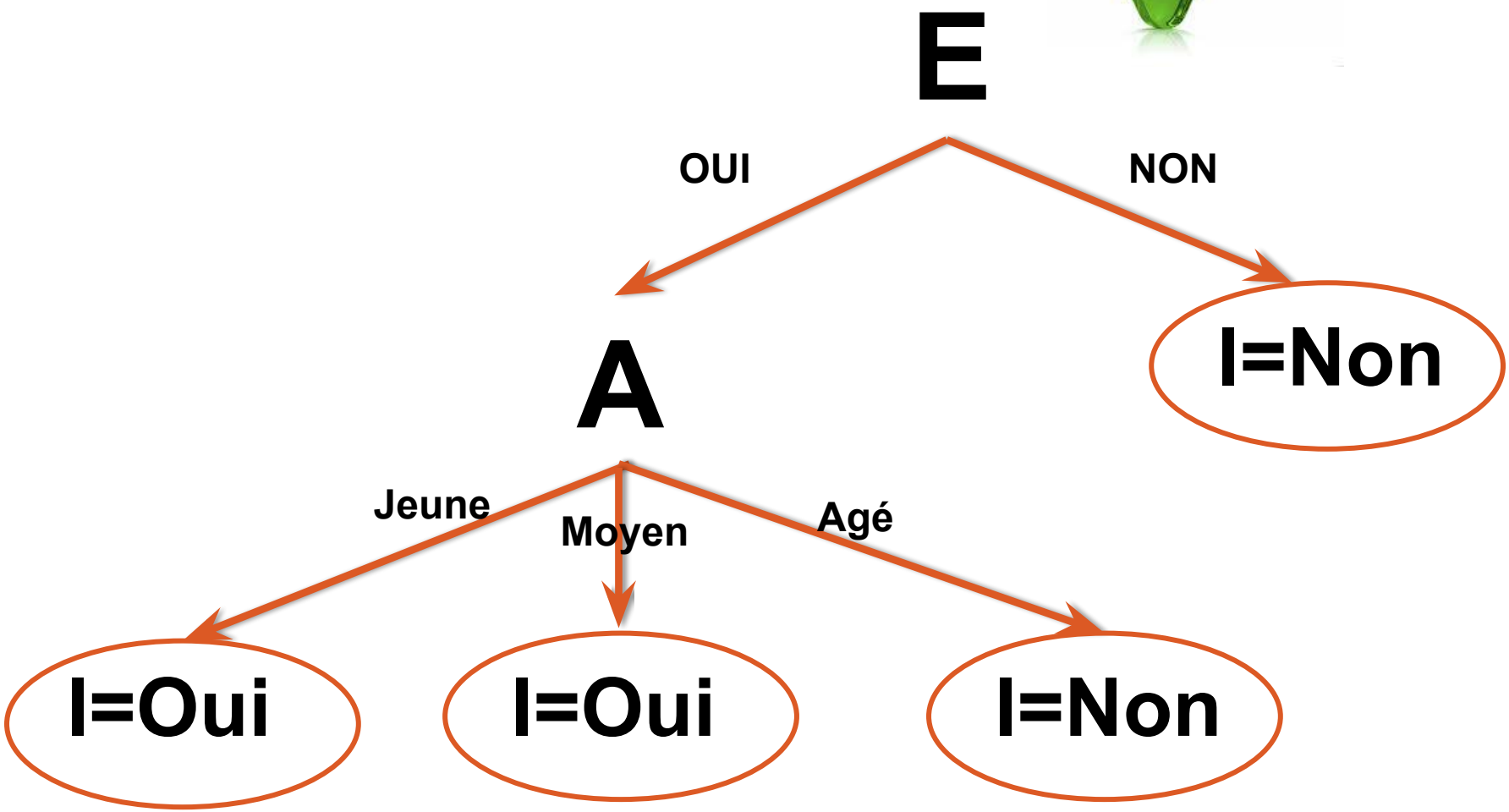
$$=$$

$$0.48 - [0 + 0 + 0.4444444]$$

$$=$$

$$0.0355556$$

L'ARBRE DE DÉCISION



EVALUATION ET VALIDATION DU MODÈLE

Validation par croisement

- Matrice de contingence
- Table de confusion
- Taux d'erreur
- Indicateur trop réducteur

Courbe ROC

Courbe LIFT

	^positif	^négatif	Total
positf	40	10	50
négatif	10	40	50
Total	50	50	100

EVALUATION ET VALIDATION DU MODÈLE

Validation par croisement

Matrice de contingence

Table de confusion

Taux d'erreur

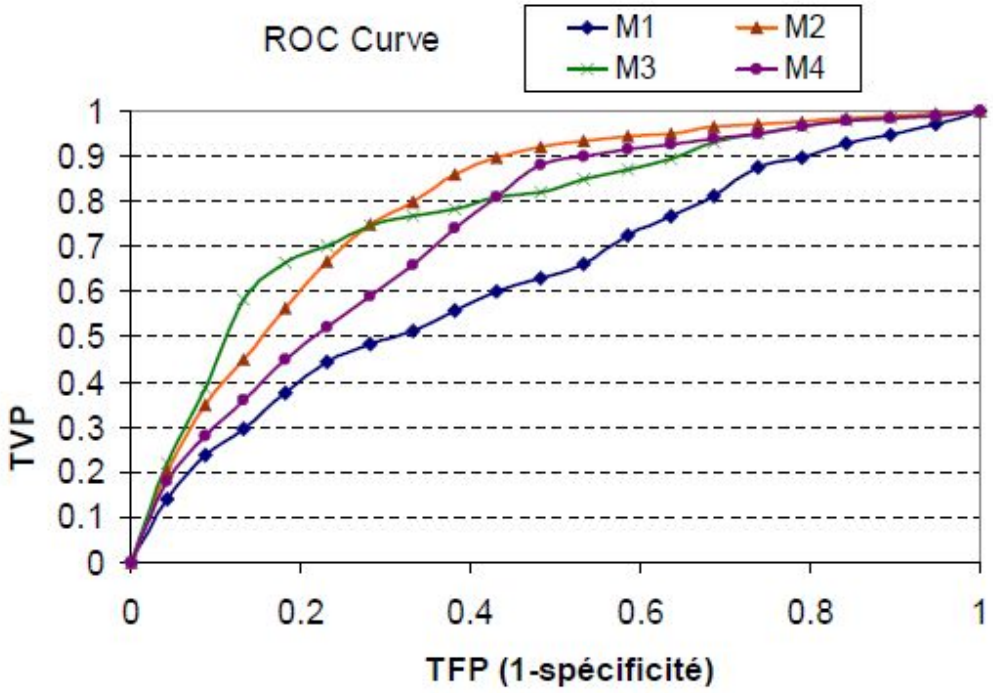
Indicateur trop réducteur

Courbe ROC

Outil d'évaluation

Outil de comparaison des modèles

Courbe LIFT



EVALUATION ET VALIDATION DU MODÈLE

Validation par croisement

Matrice de contingence

Table de confusion

Taux d'erreur

Indicateur trop réducteur

Courbe ROC

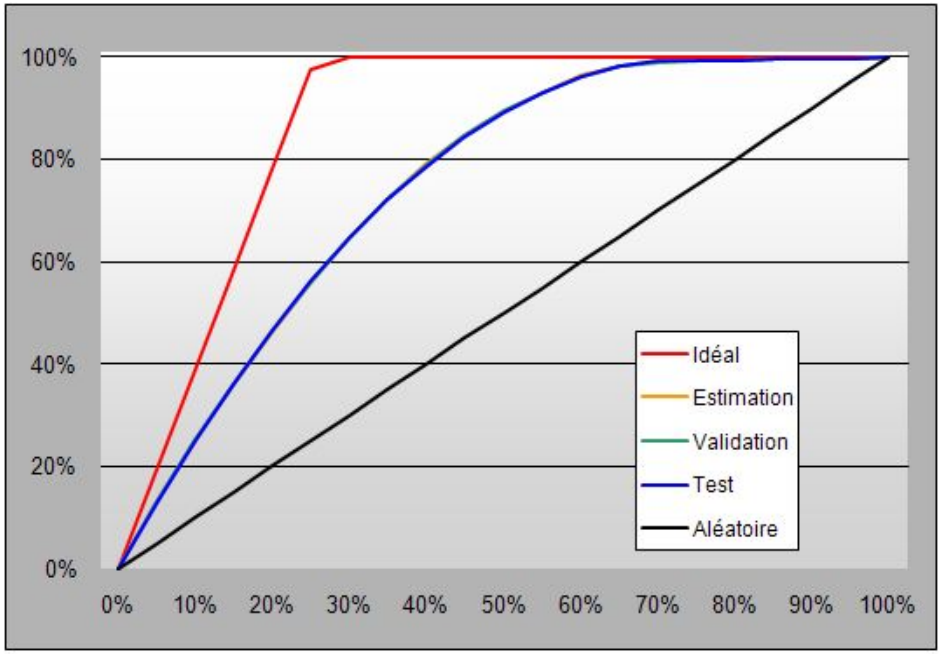
Outil d'évaluation

Outil de comparaison des modèles

Courbe LIFT

mesure de la performance d'un modèle prédictif,

Comparée au choix aléatoire



EXERCICES

esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

EXERCICE 1:

Une banque souhaite promouvoir une offre commerciale via les adresses mails de ses clients.

Pour cela elle fait appel à vous et à vos connaissances en fouille de données pour sélectionner ceux qui sont potentiellement intéressés.

Trois attributs descriptifs sont à votre disposition :

- L'âge en deux tranches : [18; 35] et [36 et plus]
- Le sexe H : Homme ou F : Femme
- Propriétaire O : oui ou N : non
- L'attribut cible qui prend deux valeurs : O (intéressé) et N (pas intéressé).

Le résultat d'une enquête préliminaire sur un échantillon représentatif de clients donne :

Age	Sexe	Propriétaire	Intéressé
20	H	N	N
25	F	N	N
32	H	O	O
34	H	O	O
37	H	N	O
41	F	O	N
45	H	O	O
45	F	O	N
52	H	O	N
60	F	O	N

EXERCICE 2:

Déduire la variable la plus décisive par rapport à l'appartenance d'un individu à l'origine orientale.

	Yeux	Cheveux	Taille	Oriental
1	Noir	Noir	Petit	Oui
2	Noir	Blanc	Grand	Oui
3	Noir	Blanc	Petit	Oui
4	Noir	Noir	Grand	Oui
5	Brun	Noir	Grand	Oui
6	Brun	Blanc	Petit	Oui
7	Bleu	Blond	Grand	Non
8	Bleu	Blond	Petit	Non
9	Bleu	Blanc	Grand	Non
10	Bleu	Noir	Petit	Non
11	Brun	Blond	Petit	Non