

TP1 TID

SCAIA Matteo, MARIAC Damien

October 24, 2024



Contents

1	Préambule	3
2	Exercice 1	3
2.1	Modélisation et variable	3
2.2	Arbre de segmentation binaire	5
3	Exercice 2	6
3.1	Recodages	6
3.1.1	Recodage en 2 variables	6
3.1.2	Recodage en 3 variables	6
3.2	Le meilleur recodage de X pour prédire Y	6
4	Exercice 3	8
4.1	Description	8
4.2	Calculs	8
4.2.1	Première étape	8
4.2.2	Deuxième étape : champignons jaune	9
4.3	Arbre de discrimination	10
5	ANNEXE	11

1 Préambule

Pour les calculs, nous avons choisit de faire l'exercice 1 (2) en utilisant le logarithme en base 2, et les exercices 2 et 3 (3 et 4) en logarithme népérien.

En effet, calculer avec le logarithme népérien ou avec le logarithme en base 2 revient au même. Les deux logarithmes sont les mêmes à un facteur près.

$$\log_2(x) = \frac{\ln(x)}{\ln(2)}$$

2 Exercice 1

On considère le tableau ci-dessous, répartissant la population active occupée selon l'âge (A), le sexe (S) et la catégorie socioprofessionnelle (C) (source: INSEE, enquête emploi 2016).

Catégorie socioprofessionnelle des actifs occupés selon le sexe et l'âge						
Âge	De 15 à 29 ans	De 30 à 49 ans	De 30 à 39 ans	De 40 à 49 ans	De 50 à 59 ans	60 ans ou plus
	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)
SEXE : Femmes						
Agriculteurs	27,8	189,7	70,0	119,6	187,1	76,9
Artisans, con	117,4	914,0	357,9	556,1	525,8	184,8
Cadres et pro	564,9	2 638,5	1 209,0	1 429,5	1 161,6	360,0
Professions i	1 353,7	3 735,7	1 840,7	1 895,0	1 507,8	256,2
Employés	1 570,9	3 486,4	1 605,6	1 880,9	1 819,6	397,0
Ouvriers	1 271,6	2 648,6	1 285,9	1 362,7	1 300,4	180,5
SEXE : Hommes						
Agriculteurs	24,2	146,0	56,2	89,8	138,0	43,4
Artisans, con	79,2	645,6	258,4	387,2	378,7	128,7
Cadres et pro	315,7	1 538,5	685,3	853,2	719,0	240,1
Professions i	613,3	1 750,4	834,7	915,7	755,9	123,7
Employés	476,0	865,5	449,5	416,0	329,8	58,3
Ouvriers	1 085,8	2 133,9	1 068,4	1 065,5	987,9	130,1

Figure 1: Tableau répartissant la population active occupée selon des catégories

2.1 Modélisation et variable

Tout d'abord de manière intuitive, nous avons envie de modéliser la variable socioprofessionnelle avec les deux autres. Cependant, nous devons le montrer de manière formelle. Grâce au code fourni dans la partie 5, nous calculons l'information mutuelle de chacune des variables.

Premièrement, nous calculons l'entropie de chacune de ces variables.

Pour la variable A , nous avons le tableau suivant (en fréquence).

	15-29 ans	30-39 ans	40-49 ans	50-59 ans	60+ ans
Age	0.1866	0.2419	0.2730	0.2441	0.0542

Table 1: Distribution par âge (A)

Nous pouvons calculer l'entropie de A .

$$H(A) = - \sum_{n=1}^6 p_i \log_2(p_i) = 2,1833$$

De même manière, nous calculons l'entropie de C et S .

	Femme	Homme
Proportion	0,6589	0,3411

Table 2: Distribution par sexe (S)

	Agriculteur	Artisans	Cadres	Profession In	Employes	Ouvrier
Proportion	0,0207	0,0740	0,1875	0,2512	0,2240	0,2423

Table 3: Distribution par catégorie socioprofessionnelle (C)

Nous obtenons.

$$H(S) = 0,9258 \quad H(C) = 2,3266$$

A présent, nous devons calculer les valeurs suivantes : $H(A, S)$, $H(A, C)$ et $H(S, C)$.

	15-29 ans	30-39 ans	40-49 ans	50-59 ans	60+
Femme	0,1220	0,1584	0,1803	0,1618	0,0362
Homme	0,0645	0,0834	0,0927	0,0823	0,1802

Table 4: Distribution jointe sexe (S) et âge (A)

	15-29 ans	30-39 ans	40-49 ans	50-59 ans	60+
Agriculteur	0,0013	0,0031	0,0052	0,0081	0,0030
Artisans	0,0049	0,0153	0,0234	0,0225	0,0080
Cadres	0,0219	0,0471	0,0568	0,0468	0,0149
Professions In	0,0489	0,0665	0,0699	0,0563	0,0094
Employes	0,0509	0,0511	0,0571	0,0535	0,0113
Ouvrier	0,0586	0,0586	0,0604	0,0569	0,0077

Table 5: Distribution jointe (C) et âge (A)

	Agriculteur	Artisans	Cadres	Profession IN	Employes	Ouvrier
Femmes	0.0119	0.0433	0.1175	0.1705	0.1810	0.1344
Homme	0.0087	0.0307	0.0700	0.0807	0.0430	0.1079

Table 6: Distribution jointe (C) et (S)

Nous obtenons les valeurs suivantes.

$$H(A, S) = 3,1092 \quad H(A, C) = 4,4817 \quad H(C, S) = 3,2242$$

De plus, nous obtenons pour $H(A, S, C)$ la valeur suivante.

$$H(A, S, C) = - \sum_{n=1}^{72} p_i \log_2(p_i) = 5,3778$$

Nous pouvons calculer les informations mutuelles.

$$I(C, (AS)) = H(C) + H(A, S) - H(A, S, C) = 0,0580$$

$$I(A, (SC)) = 0,029803$$

$$I(S, (CA)) = 0,029813$$

Cherchons le rapport entre l'information mutuelle et la variable conditionnée le plus élevé.

$$R_1 = \frac{I(C, (AS))}{H(A, C)} = 0,0187$$

$$R_2 = \frac{I(A, (SC))}{H(S, C)} = 0,0092$$

$$R_3 = \frac{I(S, (CA))}{H(A, C)} = 0,0066$$

Le rapport R_1 est le plus élevé. Donc la variable C modélisé par les deux autres (A et S) nous donne le plus d'information.

2.2 Arbre de segmentation binaire

Grâce à la partie 2.1, nous pouvons calculer les informations mutuelles suivantes.

$$I(A, S) = H(A) + H(S) - H(A, S) = 5,1104 * 10^{-5}$$

$$I(A, C) = 0,02830$$

$$I(S, C) = 0,02831$$

Faisons, la somme de ses informations mutuelles avec chacune des deux autres.

$$\bar{I}_A = I(A, S) + I(A, C) = 0,02835$$

$$\bar{I}_S = 0,02836$$

$$\bar{I}_C = 0,0566$$

Nous avons donc \bar{I}_C qui est le plus élevé. Ainsi, l'arbre de segmentation commencera avec la variable C .

3 Exercice 2

3.1 Recodages

3.1.1 Recodage en 2 variables

En agrégeant seulement les classes contigues, nous avons 4 possibilités de regroupement binaire de X .

NB : Dans cette partie, nous effectuons nos calculs avec le log népérien.

$$Z_1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$

$$Z_2 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$

$$Z_3 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$

$$Z_4 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

L'entropie de chacun de ses recodages se calcule numériquement et donne :

$$H(Z_1) = - \left(\frac{29}{72} \log_2 \left(\frac{29}{72} \right) + \frac{43}{72} \log_2 \left(\frac{43}{72} \right) \right) = 0.973$$

De la même manière, nous trouvons.

$$H(Z_2) = 0.954 \quad H(Z_3) = 0.617 \quad H(Z_4) = 0.106$$

Le meilleur recodage est donc le premiers c'est à dire : $Z_1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$

3.1.2 Recodage en 3 variables

En procédant de la même façon, on considère alors 6 cas :

$$Z_1 = \{\{0\%\}, \{0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$

$$Z_2 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$

$$Z_3 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

$$Z_4 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

$$Z_5 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$

$$Z_6 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%\}, \{> 3\%\}\}$$

L'entropie est calculé numériquement :

$$H(Z_1) = 1.541 \quad H(Z_2) = 1.462 \quad H(Z_3) = 1.068 \quad H(Z_4) = 1.040 \quad H(Z_5) = 1.320 \quad H(Z_6) = 0.684$$

Nous remarquons que le meilleure recodage en trois variables est Z_1 .

3.2 Le meilleur recodage de X pour prédire Y

Il sagit ici de recoder X en réduisant l'incertitude sur Y . Nous devons donc trouver le Z_k qui maximise l'information mutuelle entre Z_k et Y .

C'est à dire, trouvons le recodage Z_k qui maximise :

$$I(Z_k; Y) = \sum_{z \in Z_k} \sum_{y \in Y} p(z, y) \log \left(\frac{p(z, y)}{p(z)p(y)} \right) \quad (1)$$

Détaillons le calcul pour le premier recodage :

	0%	$\neq 0\%$
OUI	2/72	20/72
NON	27/72	23/72

Ainsi

$$\begin{aligned}p(Z_1 = 0\%) &= \frac{29}{72}, & p(Z_1 \neq 0\%) &= \frac{43}{72} \\p(Y = \text{Oui}) &= \frac{22}{72}, & p(Y = \text{Non}) &= \frac{50}{72}\end{aligned}$$

De plus

$$\begin{aligned}p(Z_1 = 0\%, Y = \text{Oui}) &= \frac{2}{72}, & p(Z_1 = 0\%, Y = \text{Non}) &= \frac{27}{72} \\p(Z_1 \neq 0\%, Y = \text{Oui}) &= \frac{20}{72}, & p(Z_1 \neq 0\%, Y = \text{Non}) &= \frac{23}{72}\end{aligned}$$

Nous pouvons calculer l'information mutuelle $I(Z_1, Y)$, en utilisant la formule 1.

$$\begin{aligned}I(Z_1; Y) &= p(Z_1 = 0\%, Y = \text{Oui}) \log_2 \left(\frac{p(Z_1 = 0\%, Y = \text{Oui})}{p(Z_1 = 0\%)p(Y = \text{Oui})} \right) + \dots \\&+ p(Z_1 = 0\%, Y = \text{Non}) \log_2 \left(\frac{p(Z_1 = 0\%, Y = \text{Non})}{p(Z_1 = 0\%)p(Y = \text{Non})} \right) + \dots \\&+ p(Z_1 \neq 0\%, Y = \text{Oui}) \log_2 \left(\frac{p(Z_1 \neq 0\%, Y = \text{Oui})}{p(Z_1 \neq 0\%)p(Y = \text{Oui})} \right) + \dots \\&+ p(Z_1 \neq 0\%, Y = \text{Non}) \log_2 \left(\frac{p(Z_1 \neq 0\%, Y = \text{Non})}{p(Z_1 \neq 0\%)p(Y = \text{Non})} \right)\end{aligned}$$

Nous trouvons alors:

$$I(Z_1; Y) = 0.176$$

Faisons le même cheminement mais avec les autres variables. Nous obtenons alors les informations mutuelles suivante.

$$I(Z_1, Y) = 0.176$$

$$I(Z_2, Y) = 0.091$$

$$I(Z_3, Y) = 0.033$$

$$I(Z_4, Y) = 0.024$$

On observe que l'information mutuelle la plus élevée est $I(Z_1, Y)$. Donc le recodage avec la variable Z_1 est le meilleur permettant de prédire Y .

4 Exercice 3

4.1 Description

On considère le tableau recensant des espèces de champignons suivant 4 variables qualitatives.

Table 7: Caractéristiques des espèces de champignons

Espèce	Comestible	Chapeau	Tige	Couleur
a	o	a	e	b
b	o	a	e	j
c	o	a	e	b
d	o	pl	f	j
e	o	pl	f	b
f	n	po	f	r
g	n	po	f	j
h	n	po	e	r
i	n	a	f	j
j	n	pl	f	j

On cherche à faire un arbre de discrimination qui permet de prédire la comestibilité à partir des autres caractéristiques, tout en étant le plus court possible.

Pour ce faire, on calcul les informations mutuelles de X_1 avec chaque autre variable. En effet, nous allons chercher la variables qui informe le plus sur la comestibilité. Nous effectuons nos calculs avec le logarithme népérien.

Nous posons pour la suite

$$X_1 = \{Comestible\} \quad X_2 = \{Chapeau\} \quad X_3 = \{Tige\} \quad X_4 = \{Couleur\}$$

4.2 Calculs

4.2.1 Première étape

On calcul l'information mutuelle avec la formule suivante, $I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$.

Table 8: Croisement entre comestibilité (X_1) et forme du chapeau (X_2)

$X_1 \backslash X_2$	Po	a	Pl
O	0	3	2
N	3	1	1

Nous trouvons grâce à ce tableau, $I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) = 0.277$.

Pour les variables suivantes, nous avons ces tableaux.

Table 9: Croisement entre comestibilité (X_1) et la tige (X_3)

$X_1 \backslash X_3$	e	f
O	3	2
N	1	4

Table 10: Croisement entre comestibilité (X_1) et la couleur (X_4)

$X_1 \backslash X_4$	b	j	r
O	3	2	0
N	0	3	2

De la même manière, nous pouvons calculer l'information mutuelle et nous trouvons :

$$I(X_1, X_3) = 0.106 \quad I(X_1, X_4) = 0.357$$

Nous pouvons conclure a cette étape que la variable qui apporte le plus d'information sur la comestibilité du champignon est sa couleur. En effet, l'information mutuelle $I(X_1, X_4)$ est la plus élevée.

De plus, on se rend compte dans le tableau 7 que les champignons brun sont comestibles alors que les rouges ne le sont pas.

Traitons le cas des champignons jaune.

4.2.2 Deuxieme étape : champignons jaune

En considérant que les champignons jaune, on calcul les informations mutuelles suivant les autres variables (X_2 et X_3). Nous obtenons les tableaux suivant.

Table 11: Tableau de X_1, X_2 avec champignon jaune

$X_1 \backslash X_2$	a	pl	po
O	1	1	0
N	1	1	1

Table 12: Tableau de X_1, X_3 avec champignon jaune

$X_1 \backslash X_3$	e	f
O	1	1
N	0	2

Nous calculons les informations mutuelles et nous obtenons les résultats suivant.

$$I(X_1, X_2) = 0.118 \quad I(X_1, X_3) = 0.223.$$

On considère alors la variable X_3 , celle qui correspond à la tige du champignon. Si le champignon est jaune et épais, alors il est comestible. Sinon il faut distinguer entre le champignon d (comestible) et les champignons g,i,j. Mais seul les champignons d et j partage la meme forme du chapeau (plat), alors que les champignons g et i sont respectivement; pointu, arrondi et tout deux non comestible.

4.3 Arbre de discrimination

Avec toutes les informations mutuelles calculé, on obtient alors l'arbre suivant :

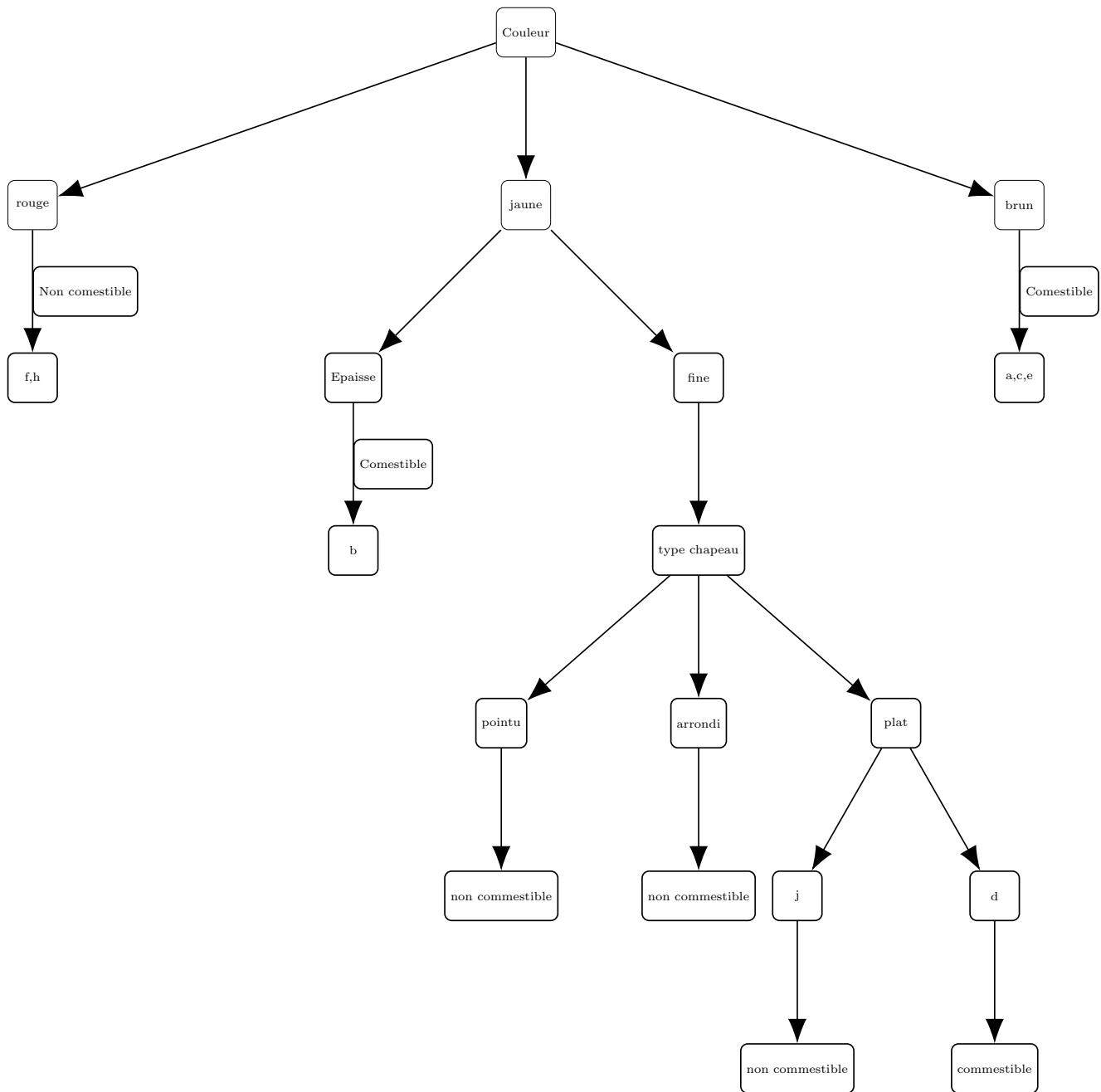


Figure 2: Arbre de discrimination pour la comestibilité d'un champignon

5 ANNEXE