TP1 TID

SCAIA Matteo, MARIAC Damien

October 26, 2024

Contents

0 Préambule 3

1 Exercice 1 3
1.1 Modélisation et variable . . . 3
1.2 Arbre de segmentation binaire . . . 5

2 Exercice 2 6
2.1 Recodages . . . 6
2.1.1 Recodage en 2 variables . . . 6
2.1.2 Recodage en 3 variables . . . 7
2.2 Le meilleur recodage de X pour prédire Y . . . 7

3 Exercice 3 9
3.1 Description . . . 9
3.2 Calculs . . . 9
3.2.1 Premiere étape . . . 9
3.2.2 Deuxieme étape : champignons jaune . . . 10
3.3 Arbre de discrimination . . . 11

4 ANNEXE 12

0 Préambule

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

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

1 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, com	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, com	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

1.1 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 4, nous calculons l'information mutuelle de chacune des variables.

Premièrement, calculons l'entropie de chacune de ces variables. Pour la variable A, nous avons le tableau suivant (en fréquence).

Table 1: Distribution par âge (A)

	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

Nous pouvons calculer l'entropie de A.

$$H(A) = -\sum_{n=1}^{5} p_i \log_2(p_i) = 2,1833$$

De la même manière, nous calculons l'entropie de C et S. Nous utilisons les tableaux ci dessous.

Table 2: Distribution par sexe (S)

	Femme	Homme
Proportion	0,6589	0,3411

Table 3: Distribution par catégorie socioprofessionnelle (C)

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

Nous obtenons.

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

A cette étape, nous pouvons interpréter les résulats. L'entropie d'une variables, mesure l'incertitude liée a celle ci. Donc la variable C est la variable avec l'incertitude la plus élevée.
Nous allons maintenant calculer les valeurs suivantes : $H(A,S)$, $H(A,C)$ et $H(S,C)$. Pour ce faire, nous déterminerons les tableaux joints correspondants pour chacune de ces paires de variables. Puis nous calculerons l'entropie des ces paires.

$$H(X) = -\sum_{x \in Tableau} P(x) \log_2(P(x))$$

Nous trouvons.

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

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

	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

Nous obtenons les valeurs suivantes.

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

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

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

Nous pouvons calculer les informations mutuelles.

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

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

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

L'information mutuelle mesure la quantité d'information partagée entre deux variables aléatoires, indiquant dans quelle mesure la connaissance de l'une des variables réduit l'incertitude concernant l'autre. Avec les calculs précédants, nous observons que la connaissance de la catégorie socioprofessionel réduit le plus l'incertitude sur l'âge et le sexe.
De plus, une information mutuelle proche de 0 sugère une indépendance.
Calculons 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. En d'autres termes, la connaissance de la catégorie socioprofesionnelle et du couple âge, sexe apporte un gain d'information sur le couple âge, catégorie socioprofesionnelle. Et ce gain est de $1,8\%$.

1.2 Arbre de segmentation binaire

Grâce à la partie 1.1, nous pouvons calculer les informations mutuelles suivantes.

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

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

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

Les résultats ci dessus, permettent de faire les interprétations suivantes. La variable âge et sexe sont proche de l'indépendance. Donc l'une n'influe presque pas sur l'autre. Et l'information mutuelle entre A et C est très proche de celle entre S et C.
Calculons maintenant la somme des informations mutuelles de chaque variable avec les deux autres. En pratique, nous souhaitons identifier la variable qui fournit le plus d'informations sur les autres. Cela revient à calculer :

$$\bar{I}_k = \sum_{n \neq k} I(X_n, X_k) \tag{1}$$

Ensuite, nous choisirons la variable pour laquelle $\bar{I}_k$ est maximale.

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

$$\bar{I_S} = 0,02836$$

$$\bar{I_C} = 0,0566$$

Nous avons donc $\bar{I_C}$ qui est le plus élevé. Ainsi, l'arbre de segmentation commencera avec la variable C, car elle fournit le plus d'information sur A et S.
Ce résultat était attendu, il correspond avec ce qu'on a trouvé dans la partie 1.1. Cela nous conforte dans l'idée que la variables C nous apporte plus d'informations sur les deux autres.
De plus, lors de la construction de l'abre, le choix de la deuxième variable de segmentations n'a pas d'inmportance, étant donné qu'il nous reste deux variables de segmentation.

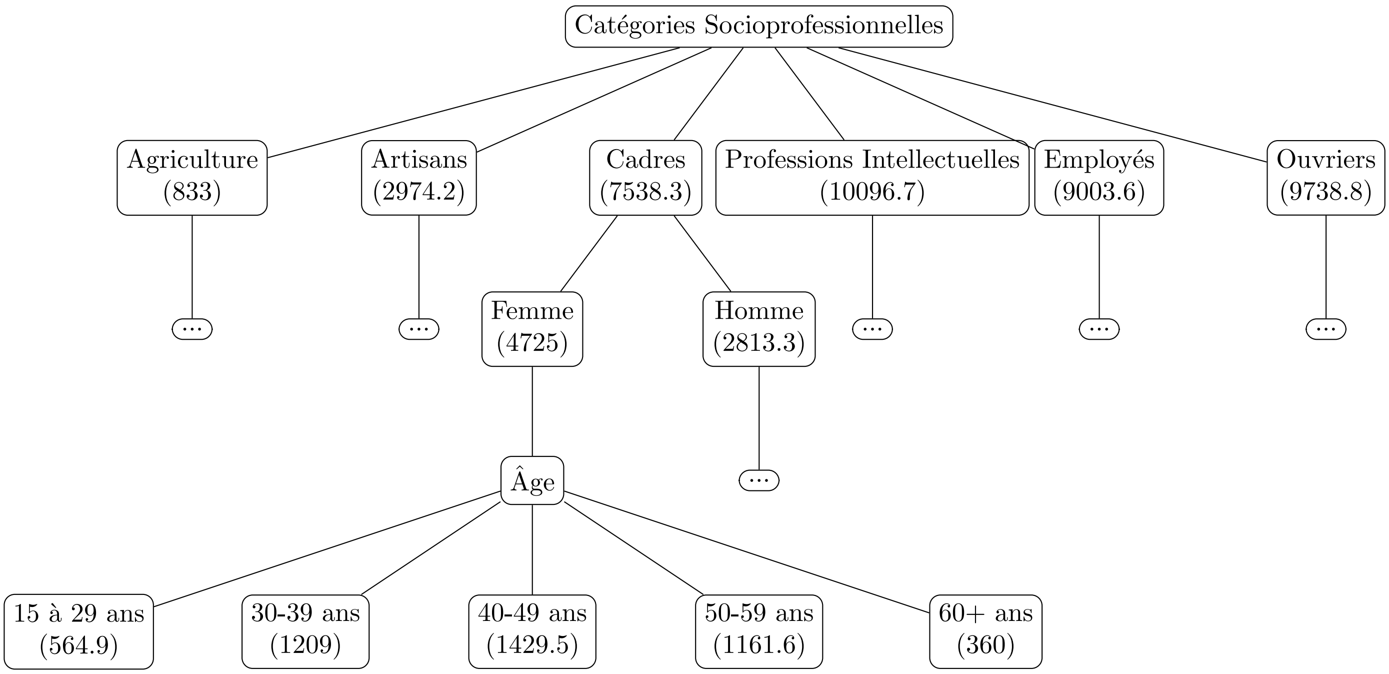

Figure 2: Arbre de segmentation binaire

2 Exercice 2

2.1 Recodages

2.1.1 Recodage en 2 variables

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

$$Z_1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$
$$Z_2 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$
$$Z_3 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$
$$Z_4 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

L'entropie de chacun de ses recodages se calcule numériquement.
Détaillons le premier calcul. Nous cherchons le tableau associé a Z_1, puis nous calculons $H(Z_1)$.

Table 7: Tableau entre Y et Z_1

$Y \backslash Z_1$	0%	{0 - 0.5 % ,0.5-1 % ,1-3 % ,¿3 % }
O	2	20
N	27	23

$$H(Z_1) = -\left(\frac{29}{72}ln(\frac{29}{72}) + \frac{43}{72}ln(\frac{43}{72})\right) = 0.674$$

De la même manière, nous trouvons.

$$H(Z_2) = 0.661 \qquad H(Z_3) = 0.427 \qquad H(Z_4) = 0.073$$

Le meilleur recodage est donc le premiers c'est à dire : $Z1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$.
En effet $H(Z_1)$ à l'entropie la plus élevée.

2.1.2 Recodage en 3 variables

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

$$\begin{aligned}
Z_1 &= \{\{0\%\}, \{0-0.5\%\}, \{0.5-1\%, 1-3\%, >3\%\}\} \\
Z_2 &= \{\{0\%\}, \{0-0.5\%, 0.5-1\%\}, \{1-3\%, >3\%\}\} \\
Z_3 &= \{\{0\%\}, \{0-0.5\%, 0.5-1\%, 1-3\%\}, \{>3\%\}\} \\
Z_4 &= \{\{0\%, 0-0.5\%\}, \{0.5-1\%, 1-3\%\}, \{>3\%\}\} \\
Z_5 &= \{\{0\%, 0-0.5\%\}, \{0.5-1\%\}, \{1-3\%, >3\%\}\} \\
Z_6 &= \{\{0\%, 0-0.5\%, 0.5-1\%\}, \{1-3\%\}, \{>3\%\}\}
\end{aligned}$$

L'entropie est calculé numériquement :

$$H(Z_1) = 1,068 \quad H(Z_2) = 1,014 \quad H(Z_3) = 0,740 \quad H(Z_4) = 0,721 \quad H(Z_5) = 0.915 \quad H(Z_6) = 0.474$$

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

2.2 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 :

$$I(Z_k; Y) = \sum_{z \in Z_k} \sum_{y \in Y} p(z, y) \log_2 \left(\frac{p(z, y)}{p(z)p(y)} \right) \tag{2}$$

Détaillons le calcul pour le premier recodage :

	0%	≠ 0%
OUI	2/72	20/72
NON	27/72	23/72

Ainsi

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

De plus

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

Nous pouvons calculer l'information mutuelle $I(Z1, Y)$, en utilisant la formule 2.

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

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 élevé est $I(Z_1, Y)$. Donc le recodage avec la variable Z_1 est le meilleur permettant de prédire Y.

3 Exercice 3

3.1 Description

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

Table 8: 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 commestibilité. Nous effecturons nos calculs avec le logarithme népérien.

Nous posons pour la suite

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

3.2 Calculs

3.2.1 Premiere 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 9: Croissement entre comestibilité (X_1) et forme du chapeau (X_2)

X1\X2	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 10: Croisement entre comestibilité (X_1) et la tige (X_3)

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

Table 11: Croisement entre comestibilité (X_1) et la couleur (X_4)

$X_1\backslash X_2$	b	j	r
O	3	2	0
N	0	3	2

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

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

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 élevé.
De plus, on se rend compte dans le tableau 8 que les champignons brun sont commestibles alors que les rouges ne le sont pas.
Traitons le cas des champignons jaune.

3.2.2 Deuxieme étape : champignons jaune

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

Table 12: 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 13: 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.

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

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

3.3 Arbre de discrimination

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

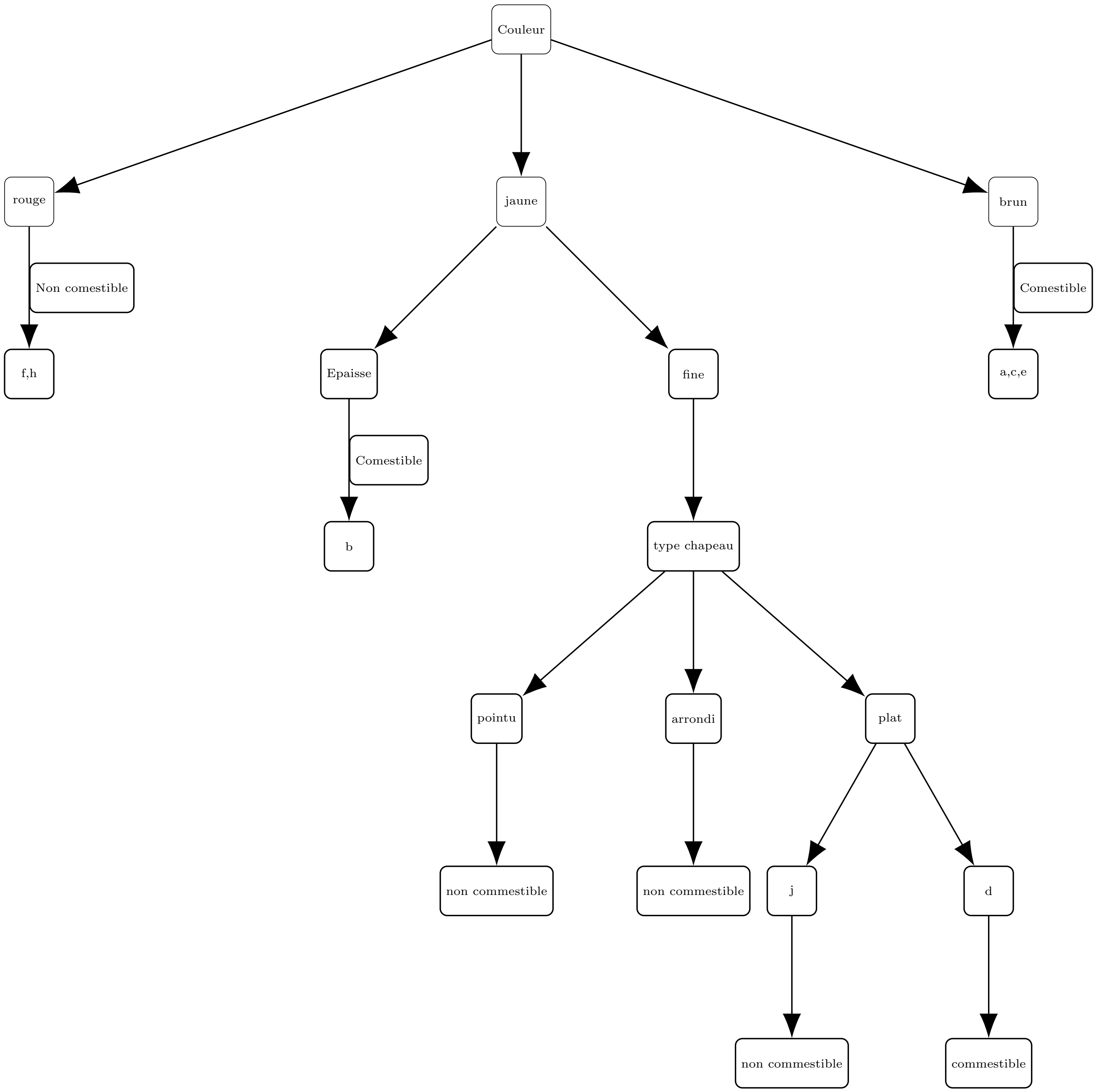

Figure 3: Arbre de discrimination pour la comestibilité d'un champignon

4 ANNEXE

```r
# Question 1
# Création d'une matrice des effectifs
effectifs <- matrix(c(
  27.8, 70.0, 119.6, 187.1, 76.9,  # Femmes - Agriculteurs
  117.4,  357.9, 556.1, 525.8, 184.8,  # Femmes - Artisans
  564.9,  1209.0, 1429.5, 1161.6, 360.0,  # Femmes - Cadres
  1353.7,  1840.7, 1895.0, 1507.8, 256.2,  # Femmes - Professions intermédiaires
  1570.9,  1605.6, 1880.9, 1819.6, 397.0,  # Femmes - Employés
  1271.6,  1285.9, 1362.7, 1300.4, 180.5,  # Femmes - Ouvriers
  24.2,  56.2, 89.8, 138.0, 43.4,  # Hommes - Agriculteurs
  79.2,  258.4, 387.2, 378.7, 128.7,  # Hommes - Artisans
  315.7,  685.3, 853.2, 719.0, 240.1,  # Hommes - Cadres
  613.3,  834.7, 915.7, 755.9, 123.7,  # Hommes - Professions intermédiaires
  476.0,  449.5, 416.0, 329.8, 58.3,  # Hommes - Employés
  1085.8,  1068.4, 1065.5, 987.9, 130.1  # Hommes - Ouvriers
), nrow = 12, byrow = TRUE)

# Noms des colonnes (groupes d age)
colnames(effectifs) <- c("15-29","30-39", "40-49", "50-59", "60+")

# Noms des lignes (combinaison sexe et catégorie socioprofessionnelle)
rownames(effectifs) <- c("Femmes_Agriculteurs", "Femmes_Artisans", "Femmes_Cadres",
                         "Femmes_Professions_inter", "Femmes_Employes", "Femmes_Ouvriers",
                         "Hommes_Agriculteurs", "Hommes_Artisans", "Hommes_Cadres",
                         "Hommes_Professions_inter", "Hommes_Employes", "Hommes_Ouvriers")

cat("Le nombre total de personne interogé", sum(effectifs))

tableau <- effectifs/sum(effectifs)  # On met notre tableau sous forme de pourcentage

#On calcul H(A)
age_effectif <- colSums(tableau)
H.A = 0
for(i in 1:length(age_effectif)){
  H.A <- H.A + (-age_effectif[i]*log2(age_effectif[i]))
}
H.A <- as.numeric(H.A)
#On calcul H(S)
femmes_effectif <- sum(rowSums(tableau[1:6,]))
hommes_effectif <- sum(rowSums(tableau[7:12,]))
sexe_effectif <- c(femmes_effectif, hommes_effectif)
H.S = 0
for(i in 1:length(sexe_effectif)){
  H.S <- H.S + (-sexe_effectif[i]*log2(sexe_effectif[i]))
}
#On calcul H(C)
# Calcul de la somme des effectifs pour chaque catégorie socioprofessionnelle
categorie_effectif <- list(
  Agriculteurs = sum(rowSums(tableau[c(1, 7), ])),       # Agriculteurs (Femmes + Hommes)
  Artisans = sum(rowSums(tableau[c(2, 8), ])),           # Artisans (Femmes + Hommes)
  Cadres = sum(rowSums(tableau[c(3, 9), ])),             # Cadres (Femmes + Hommes)
  Professions_inter = sum(rowSums(tableau[c(4, 10), ])), # Professions intermédiaires (Femmes + Hommes)
  Employes = sum(rowSums(tableau[c(5, 11), ])),          # Employés (Femmes + Hommes)
  Ouvriers = sum(rowSums(tableau[c(6, 12), ]))           # Ouvriers (Femmes + Hommes)
)
```

```
H.C = 0
for(i in 1:length(categorie_effectif)){
  H.C <- H.C + (-categorie_effectif[[i]]*log2(categorie_effectif[[i]]))
}
#Nous venons de calculer les entropies de chaque variable.

# Calculons les entropie H(A,S), H(A,C) et H(S,C)
# Pour H(A,S)
age_femme <- colSums(tableau[1:6,])
age_homme <- colSums(tableau[7:12,])
age_sexe <- c(age_femme, age_homme)
H.A.S = 0
for(i in 1:length(age_sexe)){
  H.A.S <- H.A.S + (-age_sexe[i]*log2(age_sexe[i]))
}
H.A.S <- as.numeric(H.A.S)
# Pour H(A,C)
Agriculteurs_age = colSums(tableau[c(1, 7), ])
Artisans_age = colSums(tableau[c(2, 8), ])
Cadres_age = colSums(tableau[c(3, 9), ])
Professions_inter_age = colSums(tableau[c(4, 10), ])
Employes_age = colSums(tableau[c(5, 11), ])
Ouvriers_age = colSums(tableau[c(6, 12), ])
categorie_age <- list(Agriculteurs_age,Artisans_age,Cadres_age,Professions_inter_age,
    Employes_age,Ouvriers_age)
H.A.C = 0
for(i in 1:length(categorie_age)){
  for(j in 1:length(categorie_age[[1]])){
    H.A.C <- H.A.C + (-categorie_age[[i]][j]*log2(categorie_age[[i]][j]))
  }
}
H.A.C <- as.numeric(H.A.C)
#Pour H(S,C)
femmes_catesociop <- rowSums(tableau[1:6,])
hommes_catesociop <- rowSums(tableau[7:12,])
sexe_catesociop <-list(femmes_catesociop, hommes_catesociop)
H.S.C = 0
for(i in 1:length(sexe_catesociop)){
  for(j in 1:length(sexe_catesociop[[1]])){
    H.S.C <- H.S.C + (-sexe_catesociop[[i]][j]*log2(sexe_catesociop[[i]][j]))
  }
}
H.S.C <- as.numeric(H.S.C)
# Calculons maintenant l'entropie suivante H(A,S,C)
H.A.S.C = 0
for (i in 1:nrow(tableau)) {
  for (j in 1:ncol(tableau)) {
    H.A.S.C <- H.A.S.C + (-tableau[i, j] * log2(tableau[i, j]))
  }
}
# Calculons les informations mutuelles I(A,SC) I(S,AC) I(C,AS)
I.A.SC = H.A + H.S.C - H.A.S.C
I.S.AC = H.S + H.A.C - H.A.S.C
I.C.AS = H.C + H.A.S - H.A.S.C
# On calcul les rapports
R1 <- I.A.SC / H.S.C
R2 <- I.S.AC / H.A.C
R3 <- I.C.AS / H.A.S

#Le rapport le plus eleve est R3, cela veut dire que l'environnement profesionnel est
    mieux expliqué par
# le sexe et l'age.
```

```
# Question 2
# En reprenant les calculs plus haut.
I.A.S = H.A + H.S - H.A.S
I.A.C = H.A + H.C - H.A.C
I.S.C = H.S + H.C - H.S.C
Ib_A = I.A.S + I.A.C
Ib_S = I.A.S + I.S.C
Ib_C = I.A.C + I.S.C
#La variable qui nous apporte le plus d'information est C
# On a donc plussieur classe de recette obtenus.
# On a Agriculteur Artisans Cadres  Professions intermédiaires Employés Ouvriers
# Maintenant que l'on choississent de faire une nouvelle segmentation avec S ou A
# Il n'y a pas de difference. Donc dans notre cas, le choix d'une deuxieme variable de
# segmentation n'a pas d'importance.
```