

# TP1 TID

SCAIA Matteo, MARIAC Damien

October 21, 2024



# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Exercice 1</b>                                  | <b>3</b>  |
| 1.1      | Modélisation et variable . . . . .                 | 3         |
| 1.2      | Arbre de segmentation binaire . . . . .            | 5         |
| <b>2</b> | <b>Exercice 2</b>                                  | <b>6</b>  |
| 2.1      | Recodages . . . . .                                | 6         |
| 2.1.1    | Recodage en 2 variables . . . . .                  | 6         |
| 2.1.2    | Recodage en 3 variables . . . . .                  | 6         |
| 2.2      | Le meilleur recodage de X pour prédire Y . . . . . | 6         |
| <b>3</b> | <b>Exercice 3</b>                                  | <b>8</b>  |
| 3.1      | Description . . . . .                              | 8         |
| 3.2      | Calculs . . . . .                                  | 8         |
| 3.2.1    | Détails pour le premier calcul . . . . .           | 8         |
| 3.3      | Informations mutuelles totales . . . . .           | 8         |
| 3.4      | Information mutuelle par forme . . . . .           | 9         |
| 3.4.1    | Chapeau de forme pointu . . . . .                  | 9         |
| 3.4.2    | Chapeau de forme arrondi . . . . .                 | 9         |
| 3.4.3    | Chapeau de forme plat . . . . .                    | 9         |
| 3.5      | Arbre de discrimination . . . . .                  | 10        |
| <b>4</b> | <b>ANNEXE</b>                                      | <b>11</b> |

# 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, 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

## 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, 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.

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

Nous pouvons calculer les informations mutuelles.

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

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

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

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 c'est la variable  $C$  modélisé par les deux autres ( $A$  et  $S$ ) qui nous donne le plus d'information.

## 1.2 Arbre de segmentation binaire

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

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

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

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

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  $C$ .

## 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.

$$Z1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$

$$Z2 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$

$$Z3 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$

$$Z4 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

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

$$H(Z1) = - \left( \frac{29}{72} \log_2 \left( \frac{29}{72} \right) + \frac{43}{72} \log_2 \left( \frac{43}{72} \right) \right) = 0.973$$

De la même manière, nous trouvons.

$$H(Z2) = 0.954 \quad H(Z3) = 0.617 \quad H(Z4) = 0.106$$

Le meilleur recodage est donc le premiers c'est à dire :  $Z1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$

#### 2.1.2 Recodage en 3 variables

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

$$Z1 = \{\{0\%\}, \{0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$

$$Z2 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$

$$Z3 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

$$Z4 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

$$Z5 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$

$$Z6 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%\}, \{> 3\%\}\}$$

L'entropie est calculé numériquement :

$$H(Z1) = 1.541 \quad H(Z2) = 1.462 \quad H(Z3) = 1.068 \quad H(Z4) = 1.040 \quad H(Z5) = 1.320 \quad H(Z6) = 0.684$$

Nous remarquons que le meilleure recodage en 3 variables est Z1.

## 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 :

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

Détaillons le calcul pour le premier recodage :

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

Ainsi

$$\begin{aligned} p(Z1 = 0\%) &= \frac{29}{72}, & p(Z1 \neq 0\%) &= \frac{43}{72} \\ p(Y = \text{Oui}) &= \frac{22}{72}, & p(Y = \text{Non}) &= \frac{50}{72} \end{aligned}$$

De plus

$$\begin{aligned} p(Z1 = 0\%, Y = \text{Oui}) &= \frac{2}{72}, & p(Z1 = 0\%, Y = \text{Non}) &= \frac{27}{72} \\ p(Z1 \neq 0\%, Y = \text{Oui}) &= \frac{20}{72}, & p(Z1 \neq 0\%, Y = \text{Non}) &= \frac{23}{72} \end{aligned}$$

Nous pouvons calculer l'information mutuelle  $I(Z1, Y)$ , en utilisant la formule 1.

$$\begin{aligned} I(Z1; Y) &= p(Z1 = 0\%, Y = \text{Oui}) \log_2 \left( \frac{p(Z1 = 0\%, Y = \text{Oui})}{p(Z1 = 0\%)p(Y = \text{Oui})} \right) + \dots \\ &+ p(Z1 = 0\%, Y = \text{Non}) \log_2 \left( \frac{p(Z1 = 0\%, Y = \text{Non})}{p(Z1 = 0\%)p(Y = \text{Non})} \right) + \dots \\ &+ p(Z1 \neq 0\%, Y = \text{Oui}) \log_2 \left( \frac{p(Z1 \neq 0\%, Y = \text{Oui})}{p(Z1 \neq 0\%)p(Y = \text{Oui})} \right) + \dots \\ &+ p(Z1 \neq 0\%, Y = \text{Non}) \log_2 \left( \frac{p(Z1 \neq 0\%, Y = \text{Non})}{p(Z1 \neq 0\%)p(Y = \text{Non})} \right) \end{aligned}$$

Nous trouvons alors.

$$I(Z1; Y) = 0.176$$

En faisant de même pour chaque information mutuelle, on trouve :

$$I(Z1, Y) = 0.176$$

$$I(Z2, Y) = 0.091$$

$$I(Z3, Y) = 0.033$$

$$I(Z4, Y) = 0.024$$

L'information mutuelle la plus grande est le recodage Z1. Cela signifie que Z1 est le meilleur recodage qui permet de prédire Y.

### 3 Exercice 3

#### 3.1 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 totales entre chaque variable. En effet en pratique, nous allons chercher la variables qui informe le plus sur les autres. Cela revient à calculer.

$$\bar{I}_k = \sum_{n \neq k=1}^4 I(X_n, X_k) \quad (2)$$

et choisir la variable tel que  $\bar{I}_k$  est maximale.

où

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

#### 3.2 Calculs

##### 3.2.1 Détails pour le premier calcul

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

**Table 8:** Croisement entre comestibilité (X1) et forme du chapeau (X2)

| X1 \ X2 | Po | a | Pl |
|---------|----|---|----|
| O       | 0  | 3 | 2  |
| N       | 3  | 1 | 1  |

Nous trouvons que  $I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) = 0.277$

En procédant de la même manière pour les cas restant, nous trouvons :

$$I(X_1, X_3) = 0.086 \quad I(X_1, X_4) = 0.357 \quad I(X_2, X_3) = 0.257 \quad I(X_2, X_4) = 0.370 \quad I(X_3, X_4) = 0.093$$

#### 3.3 Informations mutuelles totales

On calcul l'information mutuelle, avec la formule 2.

$$\bar{I}_1 = 0.720 \quad \bar{I}_2 = 0.905 \quad \bar{I}_3 = 0.437 \quad \bar{I}_4 = 0.820$$

Ainsi,  $X_2$  ("forme du chapeau") est la variable qui donne le plus d'information sur les autres.

Nous allons alors analyser les informations mutuelles en fonction de leur forme



### 3.4 Information mutuelle par forme

#### 3.4.1 Chapeau de forme pointu

On remarque qu'aucun des champignons pointu est comestible. (cf tableau 7)

#### 3.4.2 Chapeau de forme arrondi

Si on considère que les chapeau de forme arrondi, on obtient les tableaux suivant:

**Table 9:** X1 et X3 avec chapeau arrondi

| X1\X3 | E | F |
|-------|---|---|
| O     | 3 | 0 |
| N     | 0 | 1 |

**Table 10:** X1 et X4 avec chapeau arrondi

| X1\X4 | b | j | r |
|-------|---|---|---|
| O     | 2 | 1 | 0 |
| N     | 0 | 1 | 0 |

**Table 11:** X3 et X4 avec chapeau arrondi

| X3\X4 | b | j | r |
|-------|---|---|---|
| E     | 2 | 1 | 0 |
| F     | 0 | 1 | 0 |

Les information mutuelles sont alors :

$$I(X_1, X_3) = 0.562 \quad I(X_1, X_4) = 0.216 \quad I(X_3, X_4) = 0.216$$

On a alors :

$$\bar{I}_1 = I(X_1, X_3) + I(X_1, X_4) = 0.778$$

$$\bar{I}_3 = I(X_1, X_3) + I(X_3, X_4) = 0.778$$

$$\bar{I}_4 = I(X_1, X_4) + I(X_3, X_4) = 0.432$$

On choisit pour les chapeaux arrondi la variable  $X_3$ .

#### 3.4.3 Chapeau de forme plat

Si on considère que les chapeaux de formes plat, on obtient les tableaux suivant:

**Table 12:** X1 et X3 avec chapeau plat

| X1\X3 | E | F |
|-------|---|---|
| O     | 0 | 2 |
| N     | 0 | 1 |

**Table 13:** X1 et X4 avec chapeau plat

| X1\X4 | b | j | r |
|-------|---|---|---|
| O     | 1 | 1 | 0 |
| N     | 0 | 1 | 0 |

**Table 14:** X3 et X4 avec chapeau plat

| X3\X4 | b | j | r |
|-------|---|---|---|
| E     | 0 | 1 | 0 |
| F     | 1 | 1 | 0 |

Les information mutuelles sont alors :

$$I(X_1, X_3) = 0;$$

$$I(X_1, X_4) = 0.174;$$

$$I(X_3, X_4) = 0.174;$$

Donc on obtient :

$$\bar{I}_1 = 0.174$$

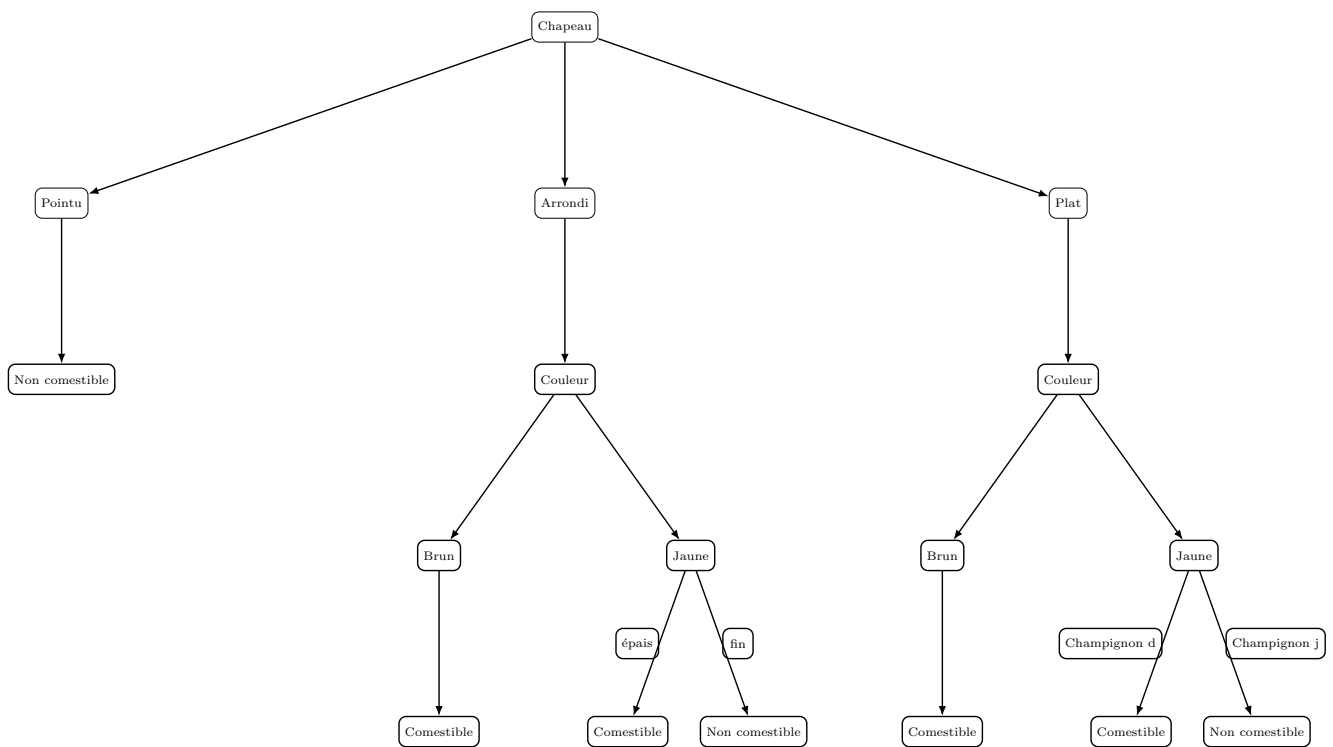
$$\bar{I}_3 = 0.174$$

$$\bar{I}_4 = 0,348$$

### 3.5 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



## 4 ANNEXE