Systèmes d'Information Décisionnels
RICM5

- Durée : 2h heures
- Documents autorisés.
- Des réponses brèves, mais claires, sont attendues.

Problème : Analyse de données d'assurances Nous disposons d'une base de données « Assurance » donnant la description de 64 contrats d'assurance voiture par cinq variables [1] :

- ***District*** : une variable nominale codant les régions de 1 à 4,
- ***Age*** : une variable ordinale indiquant la tranche d'âge des assurés de modalités : <25, 25–29, 30–35, >35,
- ***Holders*** : une variable continue indiquant le nombre d'assurés
- ***Group*** : une variable ordinale indiquant la catégorie des voitures assurées de modalités : <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre,
- ***Claim*** : une variable continue correspondant au nombre de réclamations effectuées.

Les données « Assurance » sont illustrées par l'extrait de la base suivant :

Row nb	District	Group	Age	Holders	Claims
1	1	<1l	<25	197	38
2	1	1.5l	25-29	536	84
17	2	<1l	<25	85	22
18	2	1-1.5l	>35	2443	290
34	3	<1l	25-29	73	73
35	3	<1l	30-35	89	10
60	4	1.5-2l	>35	344	63
61	4	>2l	<25	3	0

Partie 1 : Analyse descriptive

Considérons les commandes suivantes et les résultats d'exécution correspondant :

```
Ins<-Assurance
attributes(Ins)
summary(Ins)
```

[1] L. A. Baxter, S. M. Coutts and G. A. F. Ross (1980) Applications of linear models in motor insurance. *Proceedings of the 21st International Congress of Actuaries, Zurich* pp. 11–29.

```
$names
[1] "District" "Group"    "Age"      "Holders"  "Claims"
$class
[1] "data.frame"
$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 54 55 56 57 58 59 60 61 62 63 64
> dim(Ins)
[1] 64  5
```

District	Group	Age	Holders	Claims
1:16	<1l:16	<25 :16	Min. : 3.00	Min. : 0.00
2:16	1-1.5l:16	25-29:16	1st Qu. : 46.75	1st Qu.: 9.50
3:16	1.5-2l:16	30-35:16	Median : 136.00	Median : 22.00
4:16	>2l:16	>35 :16	Mean : 364.98	Mean : 49.23
			3rd Qu. : 327.50	3rd Qu.: 55.50
			Max. : 3582.00	Max. : 400.00

a) Interprétez les résultats correspondants aux variables « District » et « Holders »
b) Que peut-on dire de la distribution du nombre de contrats d'assurance conclus par région, par catégorie de voiture et par âge ?

Considérons le script suivant ainsi que les résultats associés :

```
par(mfrow=c(1,4))
boxplot(split(Ins$Group,Ins$District),main="Group")
boxplot(split(Ins$Age,Ins$District),main="Age")
boxplot(split(Ins$Holders,Ins$District),main="Holders")
boxplot(split(Ins$Claims,Ins$District),main="Claims")
```

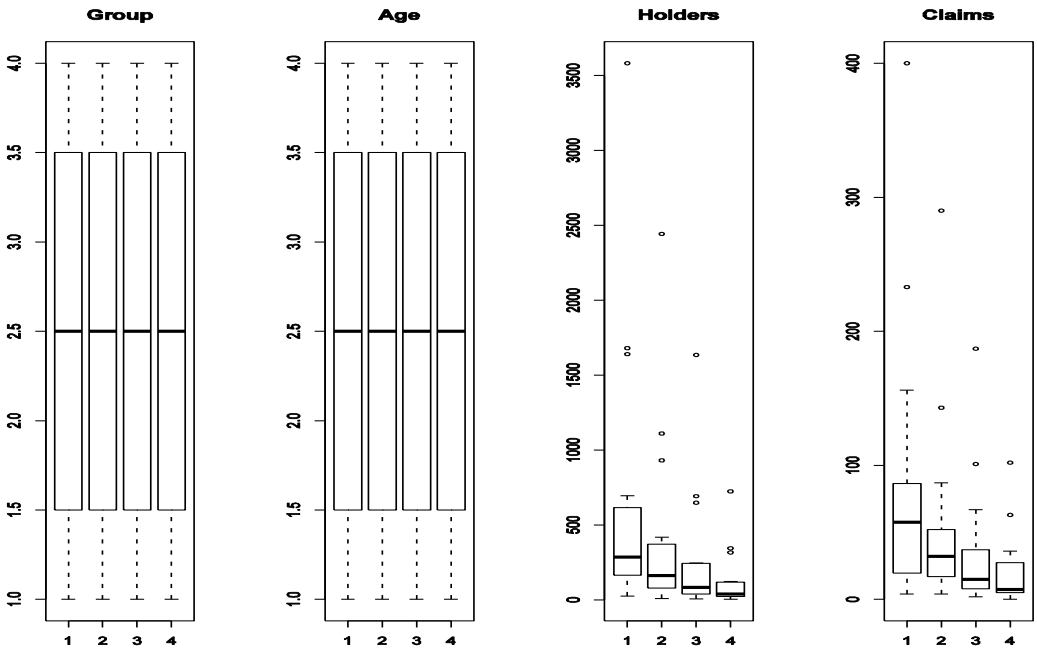


Figure 1.

c) Interprétez les boxplots de la Figure 1.

Partie 2 : Classification non-supervisée des données d'assurance.
Notre objectif est de procéder à une classification non supervisée des 64 observations via deux approches : pam, puis confronter les résultats obtenus avec ceux d'une classification hiérarchique. Pour cela, le script suivant est exécuté :

```
1.  par(mfrow=c(1,3))
2. DistIns<-daisy(Ins[,c(-1,-2)])
3. si<-c(2:20)
4.    for(nbc in 2:20){
5.      resupam<-pam( DistIns,k=nbc)
6.      si[nbc-1]<- resupam$silinfo$avg.width
7.    }
8.    nbc<-c(2:20)
9.    plot(nbc,si, type="l", main="Evolution de la valeur silhouette")
10  N<-which(si==max(si))+1
11.  resupam<-pam(DistIns,k=N)
12.  plot(resupam)
13. resuhclust <-hclust(DistIns)
14.  plot(resuhclust, hang=-1)
```

a) Expliquez clairement l'intérêt de l'instruction à la ligne 2, des instructions entre la ligne 4 et 7, et l'instruction 11.
b) Interprétez les deux graphiques donnés en Figures 2, 3, et 4.

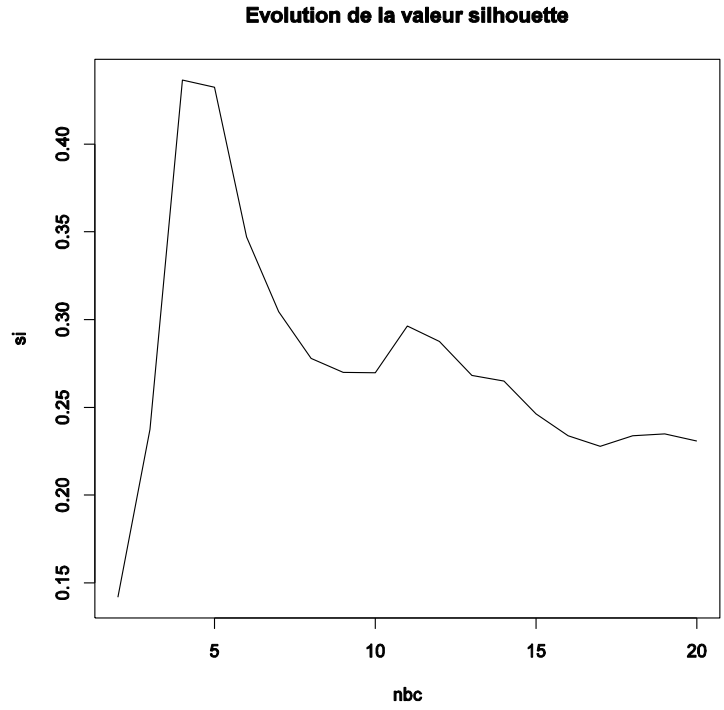


Figure 2.

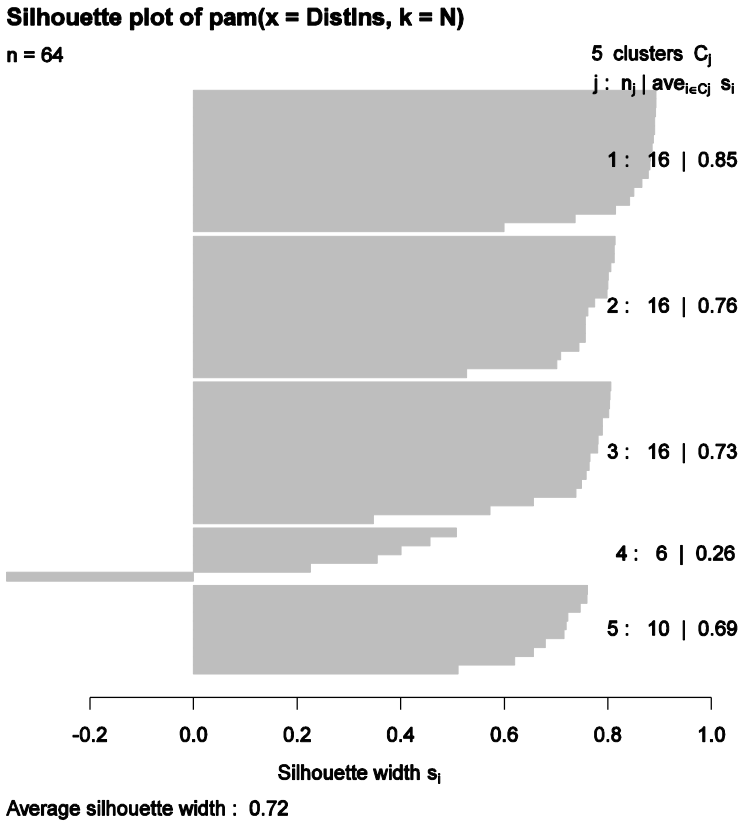


Figure 3.

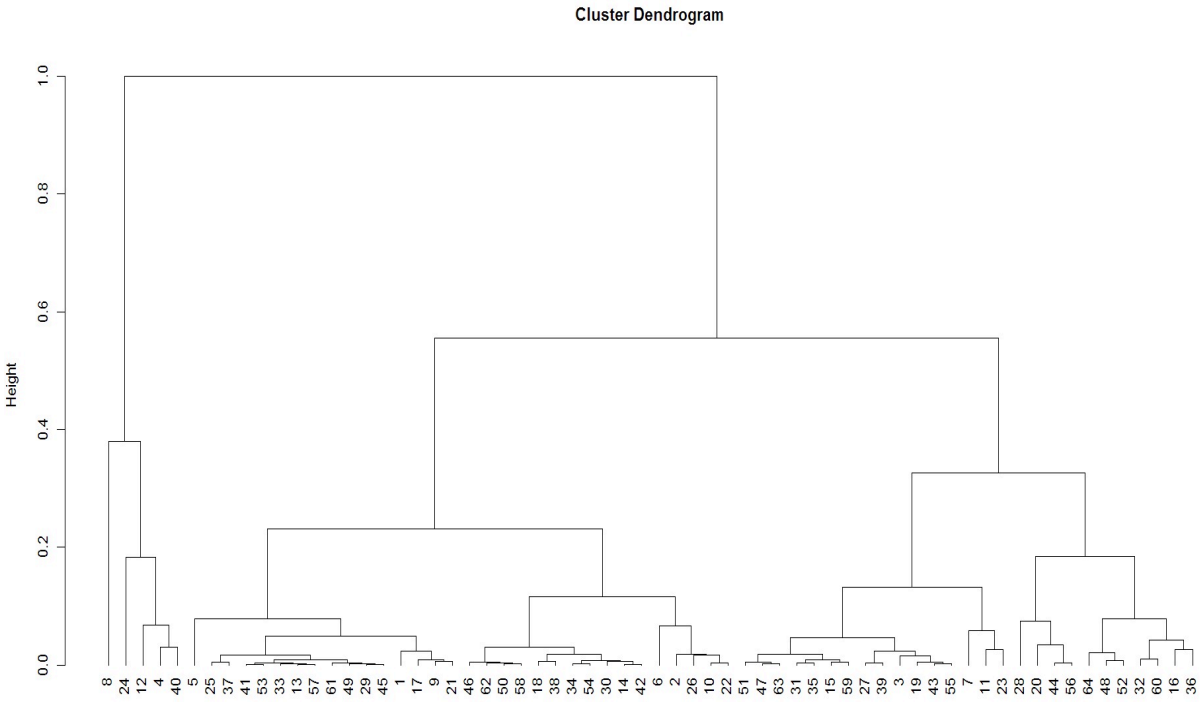


Figure 4.

c) Quel conclusion en tirer quand au bon nombre de classe.

Nous visualisons le profil des individus d'identifiant 8,24, 12, 4, et 40

	District	Group	Age	Holders	Claims
8	1	1-1.5l	>35	3582	400
24	2	1-1.5l	>35	2443	290
12	1	1.5-2l	>35	1640	233
4	1	<1l	>35	1680	156
40	3	1-1.5l	>35	1635	187

d) Quel interprétation donner à cette classe.

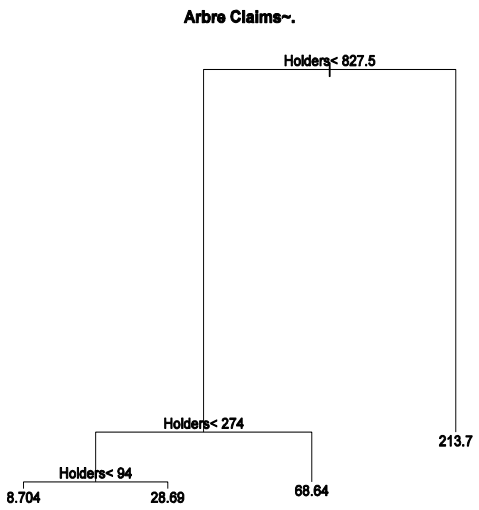
Partie 3. Classification supervisée.

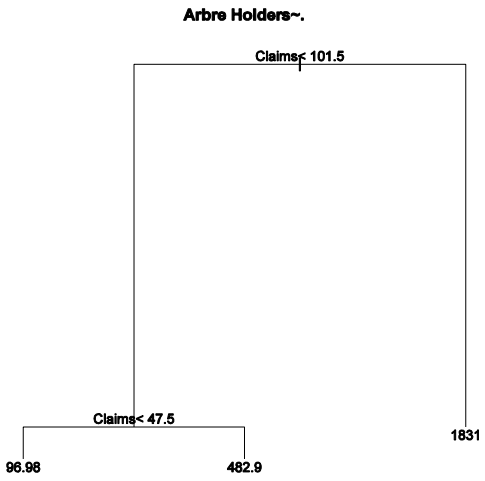
Nous procédons à la classification par arbre des données d'assurance afin d'explorer d'éventuelles règles régissant les données d'assurance. On utilise le script suivant :

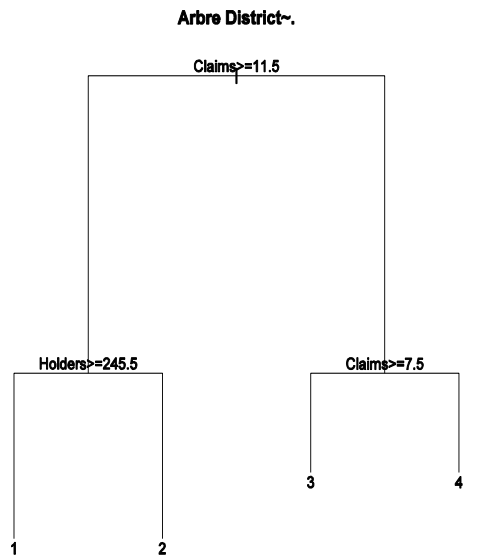
```
par(mfrow=c(1,1))
resutreeC<-rpart(Ins$Claims~., data=Ins)
plot(resutreeC)
text(resutreeC)
resutreeH<-rpart(Ins$Holders~., data=Ins)
plot(resutreeH)
text(resutreeH)
resutreeD<-rpart(Ins$District~., data=Ins)
plot(resutreeD)
text(resutreeD)
```

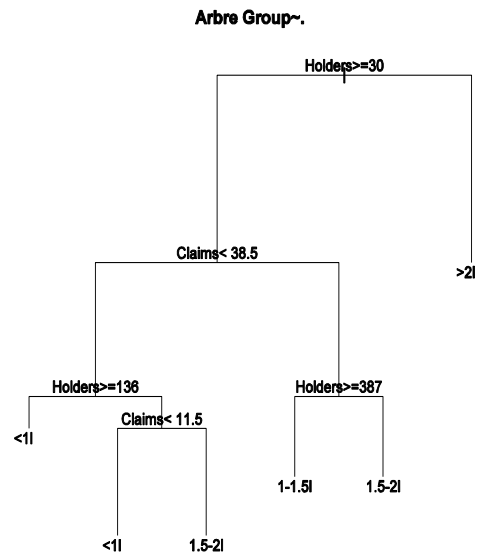
```
resutreeG<-rpart(Ins$Group~., data=Ins)
plot(resutreeG)
text(resutreeG)
par(mfrow=c(1,1))
resutreeA<-rpart(Ins$Age~., data=Ins)
plot(resutreeA, angle=-90)
text(resutreeA)
```

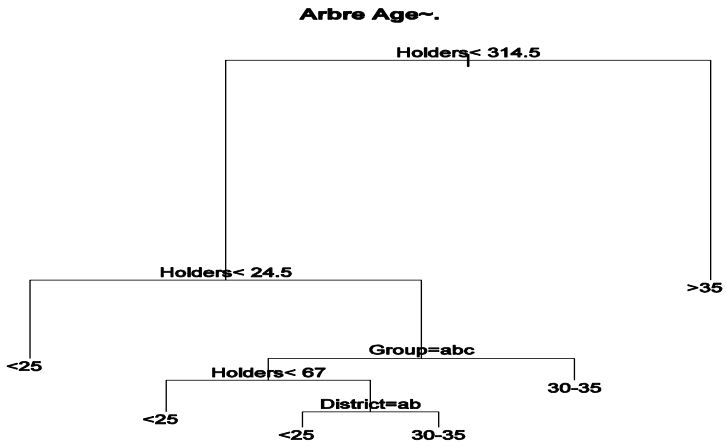
a) Lequel des arbres construits minimise l'erreur apparente de classement ou de régression ?
b) Donnez la liste des règles de décision du meilleur arbre trouvé en a).
c) Quelle(s) critique(s) méthodologique(s) pouvez-vous formuler concernant la classification par arbre effectuée ? Argumentez.











Arbre Claims~.

```
n= 64

node), split, n, deviance, yval
      * denotes terminal node

1) root 64 319037.5000  49.234380
  2) Holders< 827.5 57  40409.9300  29.035090
    4) Holders< 274 43   6037.1630  16.139530
      8) Holders< 94 27    831.6296   8.703704 *
      9) Holders>=94 16   1193.4380  28.687500 *
    5) Holders>=274 14   5259.2140  68.642860 *
  3) Holders>=827.5 7  65995.4300 213.714300 *
```

Arbre Holder~.

```
n= 64

node), split, n, deviance, yval
      * denotes terminal node

1) root 64 24434120.0  364.98440
  2) Claims< 101.5 57  2268494.0  185.00000
    4) Claims< 47.5 44   279945.0   96.97727 *
    5) Claims>=47.5 13   493780.9  482.92310 *
  3) Claims>=101.5 7  5283532.0 1830.57100 *
```

Arbre District~.

```
n= 64

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 64 48 1 (0.25000000 0.25000000 0.25000000 0.25000000)
  2) Claims>=11.5 44 29 1 (0.34090909 0.34090909 0.18181818 0.13636364)
    4) Holders>=245.5 23 12 1 (0.47826087 0.26086957 0.13043478 0.13043478) *
    5) Holders< 245.5 21 12 2 (0.19047619 0.42857143 0.23809524 0.14285714) *
  3) Claims< 11.5 20 10 4 (0.05000000 0.05000000 0.40000000 0.50000000)
    6) Claims>=7.5 7  2 3 (0.00000000 0.00000000 0.71428571 0.28571429) *
    7) Claims< 7.5 13  5 4 (0.07692308 0.07692308 0.23076923 0.61538462) *
```

Arbre Group~.

```
n= 64

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 64 48 <1l (0.25000000 0.25000000 0.25000000 0.25000000)
   2) Holders>=30 54 38 1-1.5l (0.27777778 0.29629630 0.25925926 0.16666667)
     4) Claims< 38.5 32 20 <1l (0.37500000 0.21875000 0.18750000 0.21875000)
       8) Holders>=136 10  4 <1l (0.60000000 0.30000000 0.00000000 0.10000000) *
       9) Holders< 136 22 16 <1l (0.27272727 0.18181818 0.27272727 0.27272727)
        18) Claims< 11.5 10  5 <1l (0.50000000 0.30000000 0.10000000 0.10000000) *
        19) Claims>=11.5 12  7 1.5-2l (0.08333333 0.08333333 0.41666667 0.41666667) *
     5) Claims>=38.5 22 13 1-1.5l (0.13636364 0.40909091 0.36363636 0.09090909)
      10) Holders>=387 14  7 1-1.5l (0.21428571 0.50000000 0.21428571 0.07142857) *
      11) Holders< 387 8  3 1.5-2l (0.00000000 0.25000000 0.62500000 0.12500000) *
   3) Holders< 30 10  3 >2l (0.10000000 0.00000000 0.20000000 0.70000000) *
```

Arbre Age~.

```
n= 64

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 64 48 <25 (0.25000000 0.25000000 0.25000000 0.25000000)
   2) Holders< 314.5 46 30 <25 (0.34782609 0.32608696 0.28260870 0.04347826)
     4) Holders< 24.5 8  1 <25 (0.87500000 0.12500000 0.00000000 0.00000000) *
     5) Holders>=24.5 38 24 25-29 (0.23684211 0.36842105 0.34210526 0.05263158)
      10) Group=<1l,1-1.5l,1.5-2l 29 18 25-29 (0.31034483 0.37931034 0.31034483 0.00000000)
        20) Holders< 67 7  3 <25 (0.57142857 0.28571429 0.14285714 0.00000000) *
        21) Holders>=67 22 13 25-29 (0.22727273 0.40909091 0.36363636 0.00000000)
          42) District=1,2 13  8 <25 (0.38461538 0.38461538 0.23076923 0.00000000) *
          43) District=3,4 9  4 30-35 (0.00000000 0.44444444 0.55555556 0.00000000) *
      11) Group=>2l 9  5 30-35 (0.00000000 0.33333333 0.44444444 0.22222222) *
   3) Holders>=314.5 18  4 >35 (0.00000000 0.05555556 0.16666667 0.77777778) *
```