Systèmes d'Information Décisionnels
RICM5

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

Analyse de données pour la prédiction de la persistance de malformations post opératoires de la cyphose dorsale.

La base de donnée « kyphosis »[1] fournit des mesures effectuées sur 81 enfants opérés d'une cyphose dorsale. Chaque individu de la base est décrit par la persistance ou pas de malformations suite à l'intervention chirurgicale (kyphosis), son age en mois (Age), le nombre de vertèbres malformées (Number), et le numéro de la première vertèbre opérée en partant du haut (Start).

A) Analyse descriptive des données « kyphosis »

Pour une analyse descriptive des données «kyphosis», observez les résultats suivants :

```
> kyphosis[1:5,]
  Kyphosis Age Number Start
1   absent  71      3     5
2   absent 158      3    14
3   present 128     4     5
4   absent   2      5     1
5   absent   1      4    15

> summary(kyphosis)
   Kyphosis       Age              Number          Start
    absent :64   Min.   :  1.00   Min.   : 2.000   Min.   : 1.00
    present:17   1st Qu.: 26.00   1st Qu.  : 3.000   1st Qu.  : 9.00
                 Median : 87.00   Median : 4.000   Median :13.00
                 Mean   : 83.65   Mean   : 4.049   Mean   :11.49
                 3rd Qu.:130.00   3rd Qu. : 5.000   3rd Qu.  :16.00
                 Max.   :206.00   Max.    :10.000   Max.    :18.00
                 Median :13.00    Mean   :11.49     3rd Qu.  :16.00
```

[1] John M. Chambers and Trevor J. Hastie eds. (1992) *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, CA 1992

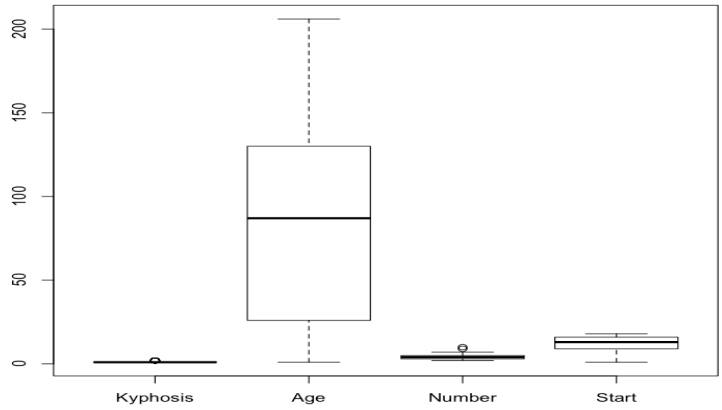

Fig. 1

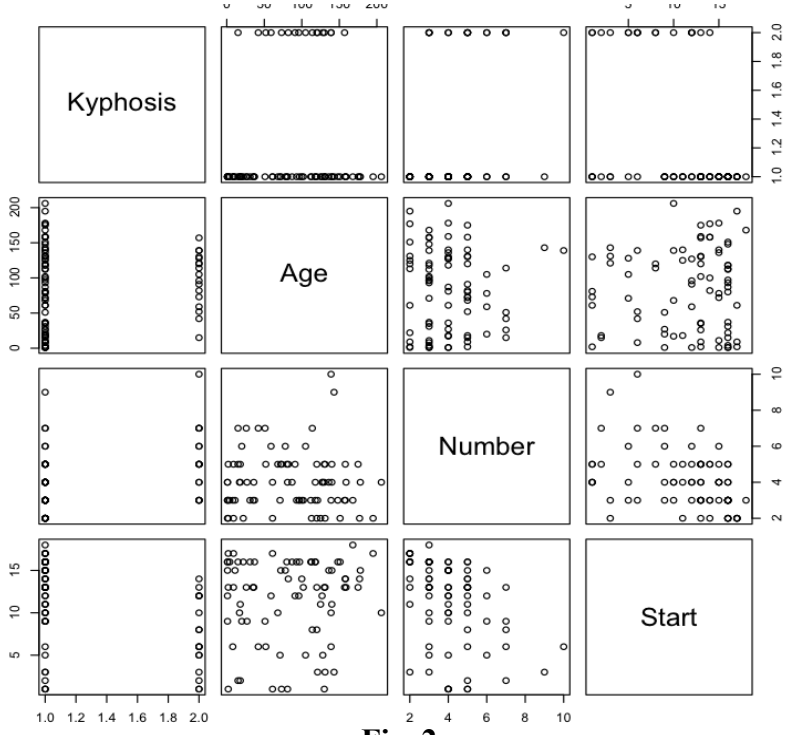

Fig. 2

a) Précisez la nature des descripteurs.
b) Indiquez l'age moyen des enfants atteints d'une cyphose dorsale ?
c) Quel nombre de vertèbres touchées est observé chez 75% des enfants ?
d) Commentez la distribution des mesures Age, Number et Start (Fig. 1).
e) Sur la base des projections de la Fig.2, discutez la nature des dépendances entre les couples (Age, Number), (Age, Start), et (Number, Start).

B) Classification non supervisée

Dans cette section, on utilise la méthode *k*-means pour le partitionnement données « kyphosis » pour plusieurs valeurs du nombre de classes. La progression de la variance intra-classe des différentes partitions obtenues est indiquée en Fig. 3. Les données analysées n'incluent pas le descripteur (Kyphosis).

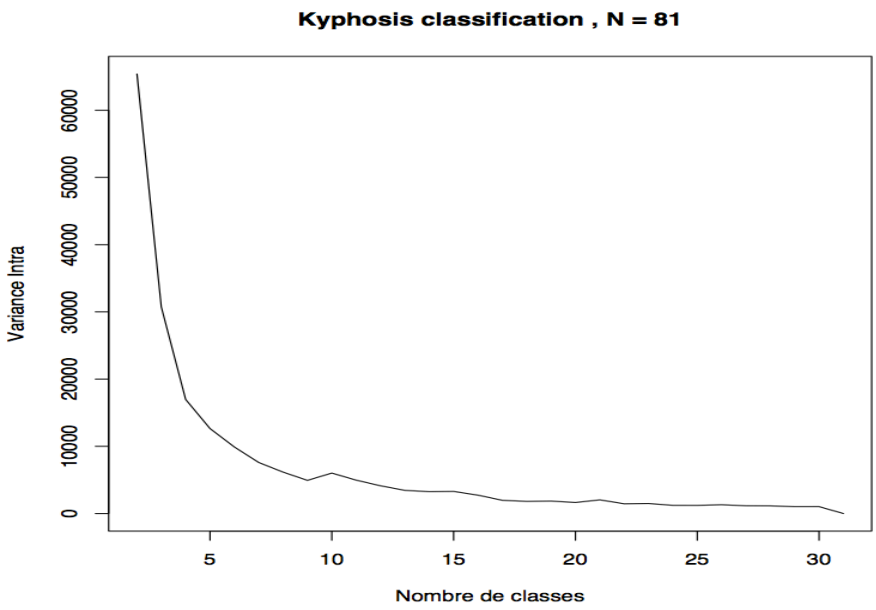

Fig.3

a) Commentez la courbe de progression de la variance intra. Expliquez le comportement observé.
b) Combien de profils de déformations post opératoires peut-on distinguer ? Justifiez.

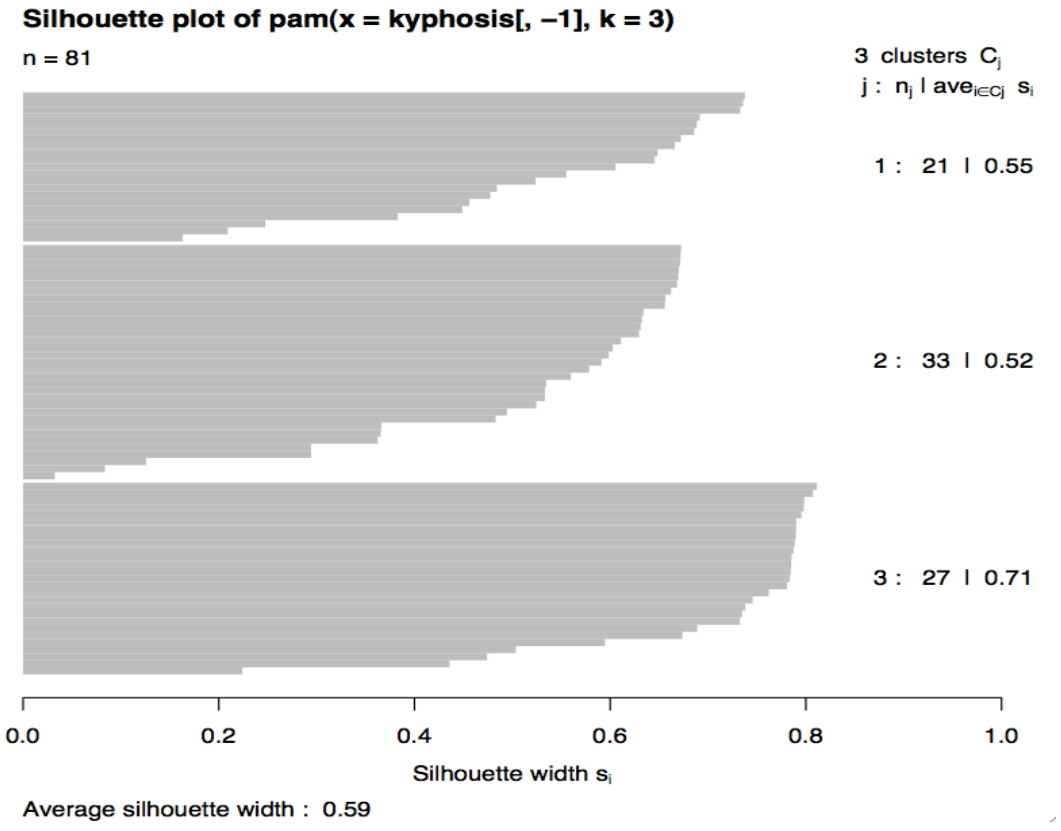

Fig. 4

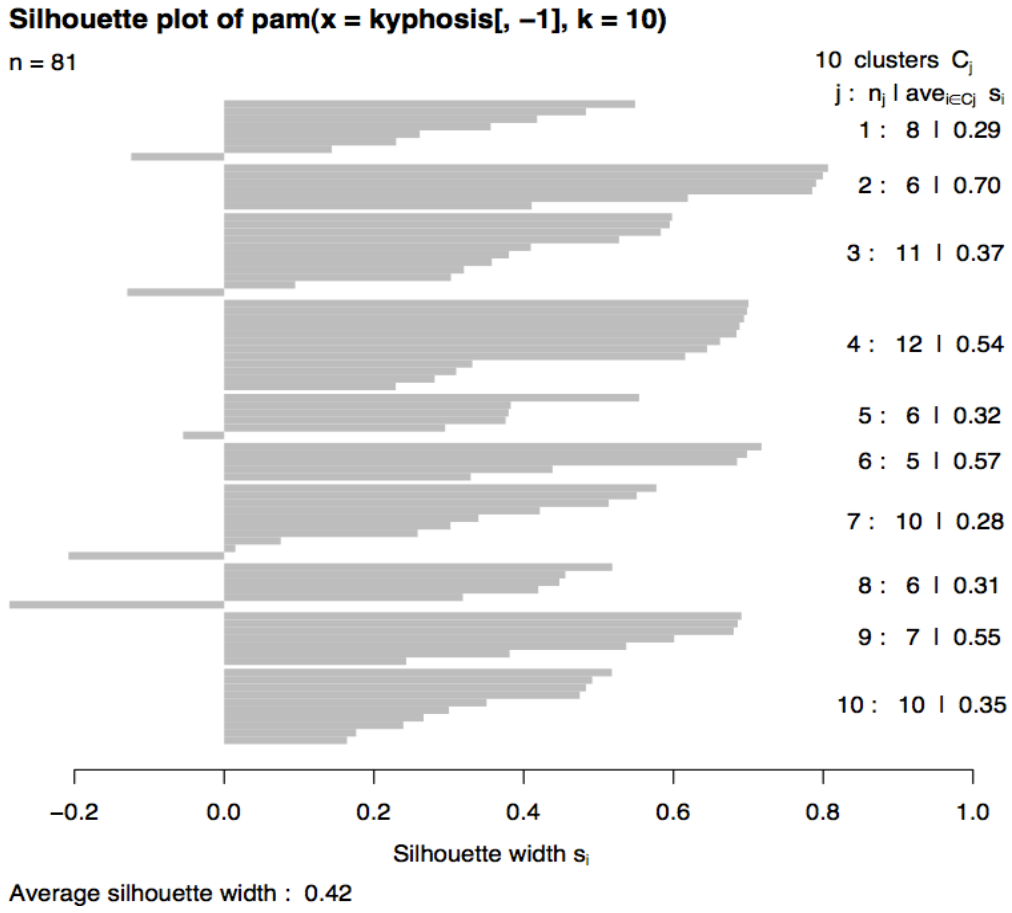

Fig. 5

L'analyse est complétée par un partitionnement PAM des données d'abord en 3, puis en 10 classes. Les figure 4 et 5 indiquent les silhouettes correspondantes.

c) Commentez ces deux figures, en particulier, indiquez les principales caractéristiques intervenant dans le choix de la meilleur partition.
d) Indiquez pour la partition de la Fig. 5 la classe la plus stable.
e) Précisez le nombre total d'individus mal classés.
f) Laquelle des deux partitions ci-dessus doit-on retenir pour une meilleure classification des données Kyphosis ?

Une troisième analyse des donnée est enfin effectuée par une classification hiérarchique ascendante dont le dendrogramme est donné en Fig. 6.

d) Interprétez le dendrogramme obtenu.
e) Sur la base des résultats des *k*-means, PAM et de la CAH, que peut-on conclure quant à la structure des profils de malformations post opératoires possibles ? Justifiez.

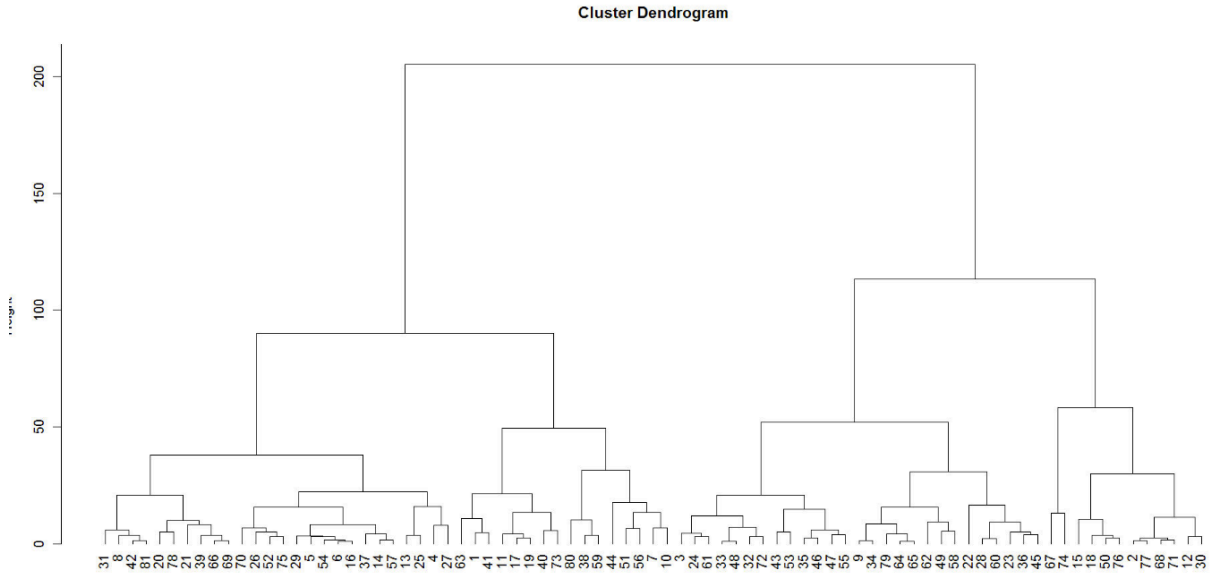

Fig. 6

C) Classification supervisée par arbre pour la prédiction de la persistance des malformations.

Notre objectif est d'expliquer la persistance de malformations de la cyphose après intervention en fonction de l'âge des enfants, le nombre de vertèbres touchées, et la localisation de la première vertèbre opérée. Pour cela, on procède à une classification par arbre selon le script suivant :

```
> fit <- rpart(Kyphosis ~ Age + Number + Start, data=kyphosis)
> plot(fit)
> text(fit, use.n=TRUE)
> fit
n= 81

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

 1) root 81 17 absent (0.79012346 0.20987654)
   2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
     4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
     5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
      10) Age< 55 12  0 absent (1.00000000 0.00000000) *
      11) Age>=55 21  6 absent (0.71428571 0.28571429)
        22) Age>=111 14  2 absent (0.85714286 0.14285714) *
        23) Age< 111 7  3 present (0.42857143 0.57142857) *
     3)  Start< 8.5 19  8 present (0.42105263 0.57894737) *
```

a) Interprétez l'arbre obtenu (Fig. 7) et en extraire des règles brèves expliquant la présence ou l'absence de malformations.

b) Quel(s) descripteur(s) se révèle(nt) indépendant(s) de la persistance de malformations.

c) Évaluez l'erreur apparente de classement de l'arbre induit.

d) Evaluez l'erreur apparente de prédiction de la persistance de malformations après intervention.

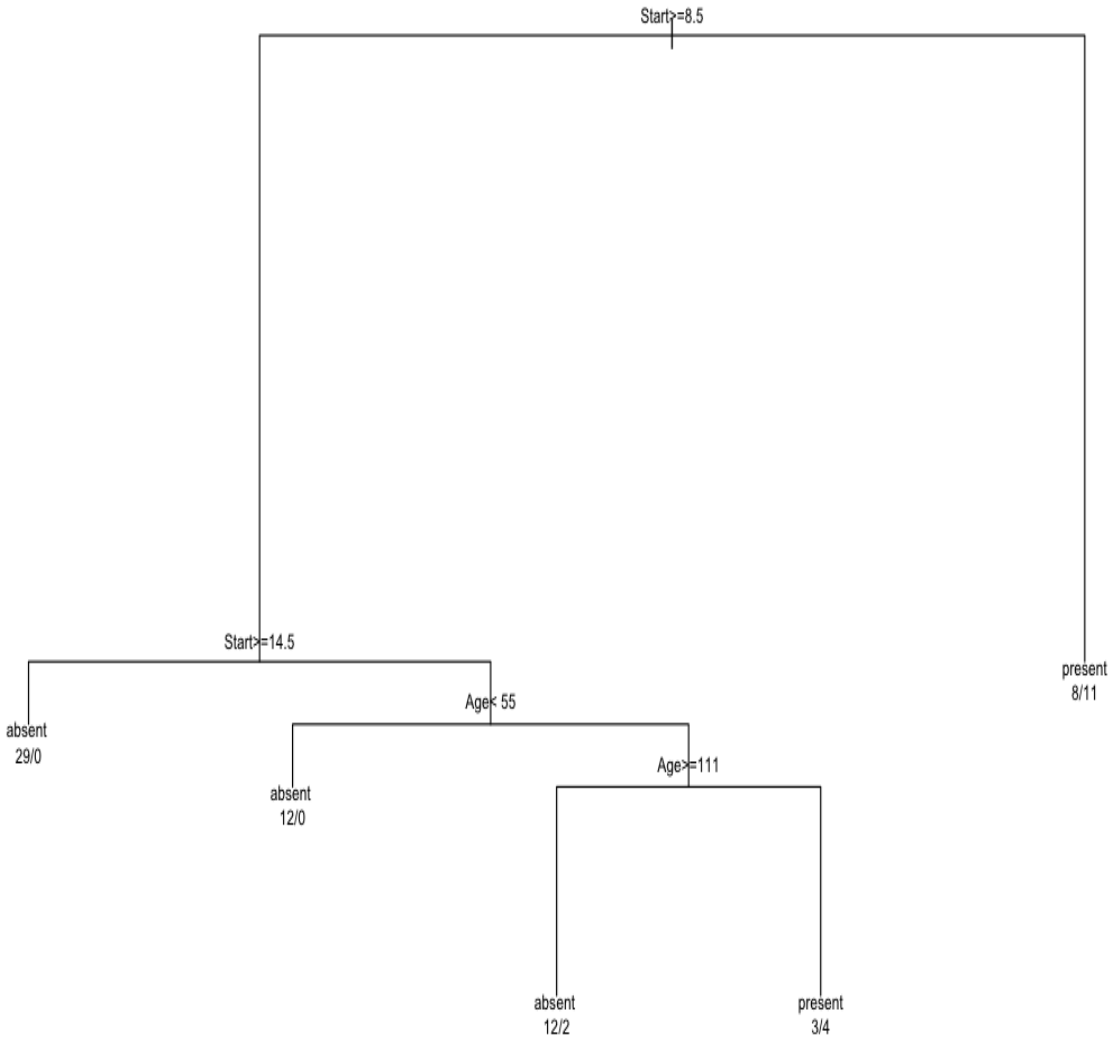
Start>=8.5
Start>=14.5
present
8/11
Age< 55
absent
29/0
absent
12/0
Age>=111
absent
12/2
present
3/4

Fig. 7