Analyse des Données NPHA

Anastasios Tsiompanidis et Noah Kohrs

2025-02-07

National Poll on Healthy Aging (NPHA)

Auteurs du Projet

Anastasios Tsiompanidis Noah Kohrs

1- Motivation et Positionnement du Projet

Ce travail est basé sur le "National Poll on Healthy Aging" (NPHA), une enquête menée auprès de personnes âgées pour évaluer leur état de santé et leurs besoins en matière de soins médicaux.

Notre analyse s'appuyera principalement sur la recherche de corrélation entre le nombre de visites chez le médecin d'un patient et ses caractéristiques exprimés par le reste des variables.

On précharge les librairies nécessaires pour l'analyse des données:

```
library(cluster)
library(rpart)
```

2- Analyse descriptive

Une première lecture des données nous donne un aperçu des valeurs uniques pour chaque variable, ce qui nous permet de détecter d'éventuels problèmes de labellisation ou de valeurs manquantes.

```
npha <- read.csv("NPHA-doctor-visits.csv")
summary(npha)
```

```
##  Number.of.Doctors.Visited      Age    Physical.Health  Mental.Health
##  Min.   :1.000             Min.   :2   Min.   :-1.000   Min.   :-1.000
##  1st Qu.:2.000             1st Qu.:2   1st Qu.: 2.000   1st Qu.: 1.000
##  Median :2.000             Median :2   Median : 3.000   Median : 2.000
##  Mean   :2.112             Mean   :2   Mean   : 2.794   Mean   : 1.989
##  3rd Qu.:3.000             3rd Qu.:2   3rd Qu.: 3.000   3rd Qu.: 3.000
##  Max.   :3.000             Max.   :2   Max.   : 5.000   Max.   : 5.000
##  Dental.Health     Employment    Stress.Keeps.Patient.from.Sleeping
##  Min.   :-1.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.: 2.00   1st Qu.:3.000   1st Qu.:0.0000
##  Median : 3.00   Median :3.000   Median :0.0000
##  Mean   : 3.01   Mean   :2.807   Mean   :0.2479
```

```
##  3rd Qu.: 4.00   3rd Qu.:3.000   3rd Qu.:0.0000
##  Max.   : 6.00   Max.   :4.000   Max.   :1.0000
##  Medication.Keeps.Patient.from.Sleeping Pain.Keeps.Patient.from.Sleeping
##  Min.   :0.00000                        Min.   :0.0000
##  1st Qu.:0.00000                        1st Qu.:0.0000
##  Median :0.00000                        Median :0.0000
##  Mean   :0.05602                        Mean   :0.2185
##  3rd Qu.:0.00000                        3rd Qu.:0.0000
##  Max.   :1.00000                        Max.   :1.0000
##  Bathroom.Needs.Keeps.Patient.from.Sleeping Unknown.Keeps.Patient.from.Sleeping
##  Min.   :0.0000                             Min.   :0.0000
##  1st Qu.:0.0000                             1st Qu.:0.0000
##  Median :1.0000                             Median :0.0000
##  Mean   :0.5042                             Mean   :0.4174
##  3rd Qu.:1.0000                             3rd Qu.:1.0000
##  Max.   :1.0000                             Max.   :1.0000
##  Trouble.Sleeping Prescription.Sleep.Medication      Race           Gender
##  Min.   :-1.000   Min.   :-1.000                Min.   :1.000   Min.   :1.00
##  1st Qu.: 2.000   1st Qu.: 3.000                1st Qu.:1.000   1st Qu.:1.00
##  Median : 3.000   Median : 3.000                Median :1.000   Median :2.00
##  Mean   : 2.408   Mean   : 2.829                Mean   :1.426   Mean   :1.55
##  3rd Qu.: 3.000   3rd Qu.: 3.000                3rd Qu.:1.000   3rd Qu.:2.00
##  Max.   : 3.000   Max.   : 3.000                Max.   :5.000   Max.   :2.00
```

Comme nos valeurs sont catégorielles représentées par des chiffres, on va les remplacer par des labels pour une meilleure compréhension.

Cependant, le dataset utilisé contient plusieurs erreurs de labellisation, ce qui nous oblige à les corriger. Nous avons essayé dee faire au plus mieux pour rester cohérent avec les valeurs existantes.

```
doctor_labels <- c("0-1", "2-3", "4+")
age_labels <- c("50-64", "65-80")

# On a ajouté la valeur "Very Poor" nous mêmes car il n'y avait
# pas de labelling indiqué pour la valeur 6.
# Cela suit la logique et nous évite la présence de NA's
health_labels <- c("Refused", "Excellent", "Very Good", "Good", "Fair", "Poor", "Very Poor")
empl_labels <- c("Refused", "Full-time", "Part-time", "Retired", "Not working")
yes_no_labels <- c("No", "Yes")
gender_labels <- c("M", "F")
medication_labels <- c("Refused", "Use regularly", "Use occasionally", "Do not use")

# Les valeurs devraient être "No" et "Yes", mais elles sont mal labellisées dans le dataset.
# Nous supposons que ces corrections sont appropriées.
sleep_labels <- c("Refused", "No", "A bit", "Yes")
race_labels <- c("Not asked", "Refused", "White", "Black", "Other", "Hispanic", "2+ Races")
```

Enfin, on utilise ces labels pour remplacer les valeurs existantes dans le dataset.

```
colnames(npha)
```

```
##  [1] "Number.of.Doctors.Visited"
##  [2] "Age"
```

```
##  [3] "Physical.Health"
##  [4] "Mental.Health"
##  [5] "Dental.Health"
##  [6] "Employment"
##  [7] "Stress.Keeps.Patient.from.Sleeping"
##  [8] "Medication.Keeps.Patient.from.Sleeping"
##  [9] "Pain.Keeps.Patient.from.Sleeping"
## [10] "Bathroom.Needs.Keeps.Patient.from.Sleeping"
## [11] "Unknown.Keeps.Patient.from.Sleeping"
## [12] "Trouble.Sleeping"
## [13] "Prescription.Sleep.Medication"
## [14] "Race"
## [15] "Gender"
```

```r
npha$Number.of.Doctors.Visited = factor(npha$Number.of.Doctors.Visited, levels = 1:3, labels = doctor_la
npha$Age = factor(npha$Age, levels = 1:2, labels = age_labels, ordered = FALSE)
npha$Physical.Health = factor(npha$Physical.Health, levels = c(-1, 1:6), labels = health_labels, ordered
npha$Mental.Health = factor(npha$Mental.Health, levels = c(-1, 1:6), labels = health_labels, ordered = 
npha$Dental.Health = factor(npha$Dental.Health, levels = c(-1, 1:6), labels = health_labels, ordered = 
npha$Employment = factor(npha$Employment, levels = c(-1, 1:4), labels = empl_labels, ordered = FALSE)
npha$Stress.Keeps.Patient.from.Sleeping = factor(npha$Stress.Keeps.Patient.from.Sleeping, levels = 0:1,
npha$Medication.Keeps.Patient.from.Sleeping = factor(npha$Medication.Keeps.Patient.from.Sleeping, level
npha$Pain.Keeps.Patient.from.Sleeping = factor(npha$Pain.Keeps.Patient.from.Sleeping, levels = 0:1, labe
npha$Bathroom.Needs.Keeps.Patient.from.Sleeping = factor(npha$Bathroom.Needs.Keeps.Patient.from.Sleeping
npha$Unknown.Keeps.Patient.from.Sleeping = factor(npha$Unknown.Keeps.Patient.from.Sleeping, levels = 0:
npha$Trouble.Sleeping = factor(npha$Trouble.Sleeping, levels = c(-1, 1:3), labels = sleep_labels, ordere
npha$Prescription.Sleep.Medication = factor(npha$Prescription.Sleep.Medication, levels = c(-1, 1:3), lab
npha$Race = factor(npha$Race, levels = 0:6, labels = race_labels, ordered = FALSE)
npha$Gender = factor(npha$Gender, levels = 1:2, labels = gender_labels, ordered = FALSE)
```

On obtient:

```r
summary(npha)
```

```
##  Number.of.Doctors.Visited    Age       Physical.Health   Mental.Health
##  0-1:131                   50-64:  0   Refused  :  1    Refused  : 10
##  2-3:372                   65-80:714   Excellent: 36    Excellent:219
##  4+ :211                               Very Good:239    Very Good:282
##                                        Good     :291    Good     :167
##                                        Fair     :126    Fair     : 34
##                                        Poor     : 21    Poor     :  2
##                                        Very Poor:  0    Very Poor:  0
##    Dental.Health       Employment  Stress.Keeps.Patient.from.Sleeping
##  Refused  :  4   Refused    :  0   No :537
##  Excellent: 66   Full-time  : 50   Yes:177
##  Very Good:215   Part-time  : 55
##  Good     :208   Retired    :592
##  Fair     :127   Not working: 17
##  Poor     : 39
##  Very Poor: 55
##  Medication.Keeps.Patient.from.Sleeping Pain.Keeps.Patient.from.Sleeping
##  No :674                                No :558
##  Yes: 40                                Yes:156
```

```
##
##
##
##
##
##  Bathroom.Needs.Keeps.Patient.from.Sleeping Unknown.Keeps.Patient.from.Sleeping
##  No :354                                    No :416
##  Yes:360                                    Yes:298
##
##
##
##
##
##  Trouble.Sleeping  Prescription.Sleep.Medication        Race     Gender
##  Refused:  2       Refused         :  3                Not asked:  0   M:321
##  No     : 62       Use regularly   : 38                Refused  :578   F:393
##  A bit  :291       Use occasionally: 34                White    : 52
##  Yes    :359       Do not use      :639                Black    : 20
##                                                       Other    : 44
##                                                       Hispanic : 20
##                                                       2+ Races :  0
```

Ce resumé est bien plus parlant et nous permet de mieux comprendre les données que nous avons à disposition.

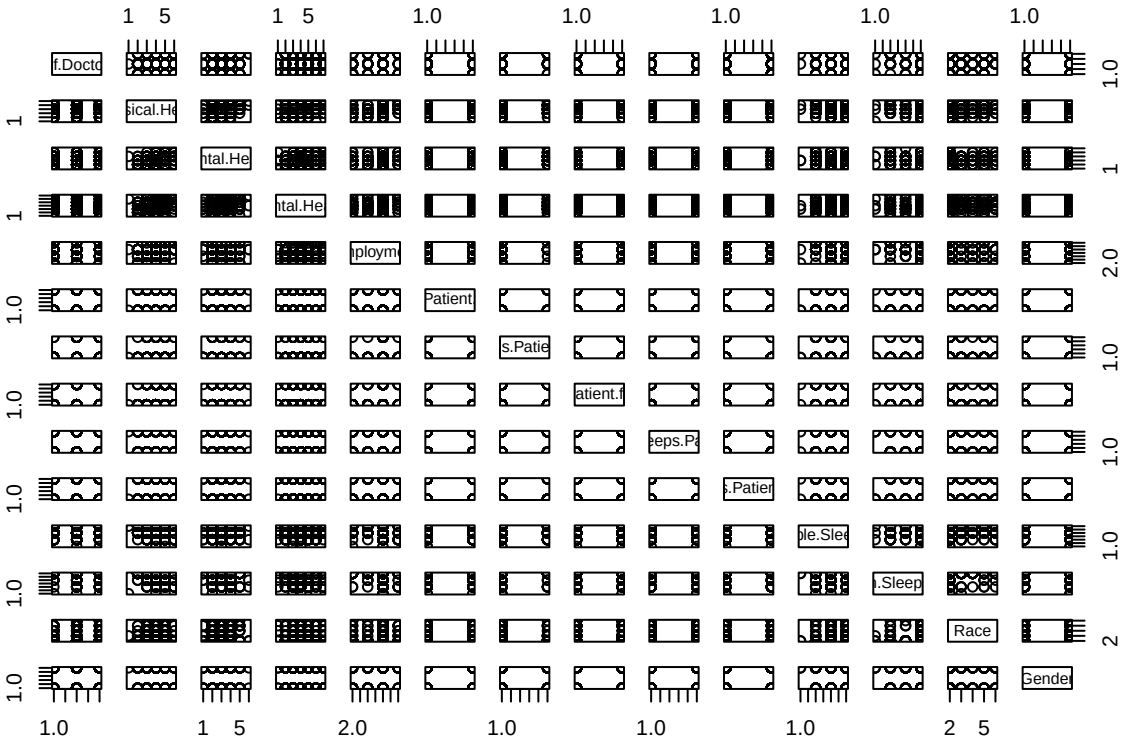
On observe par ailleurs que l'age des patients est toujours entre 65 et 80 ans, il s'agit donc d'une constante sur notre jeu de données. Nous allons donc écarter la variable de la suite de l'analyse car cela ne nous fournit aucune information utile et nuit la la lisibilité.

```
npha <- npha[, c(1, 3:ncol(npha))]
# On vérifie que l'age a bien été supprimé.
colnames(npha)
```

```
##  [1] "Number.of.Doctors.Visited"
##  [2] "Physical.Health"
##  [3] "Mental.Health"
##  [4] "Dental.Health"
##  [5] "Employment"
##  [6] "Stress.Keeps.Patient.from.Sleeping"
##  [7] "Medication.Keeps.Patient.from.Sleeping"
##  [8] "Pain.Keeps.Patient.from.Sleeping"
##  [9] "Bathroom.Needs.Keeps.Patient.from.Sleeping"
## [10] "Unknown.Keeps.Patient.from.Sleeping"
## [11] "Trouble.Sleeping"
## [12] "Prescription.Sleep.Medication"
## [13] "Race"
## [14] "Gender"
```

Essayons d'avoir une vue d'ensemble de nos données.

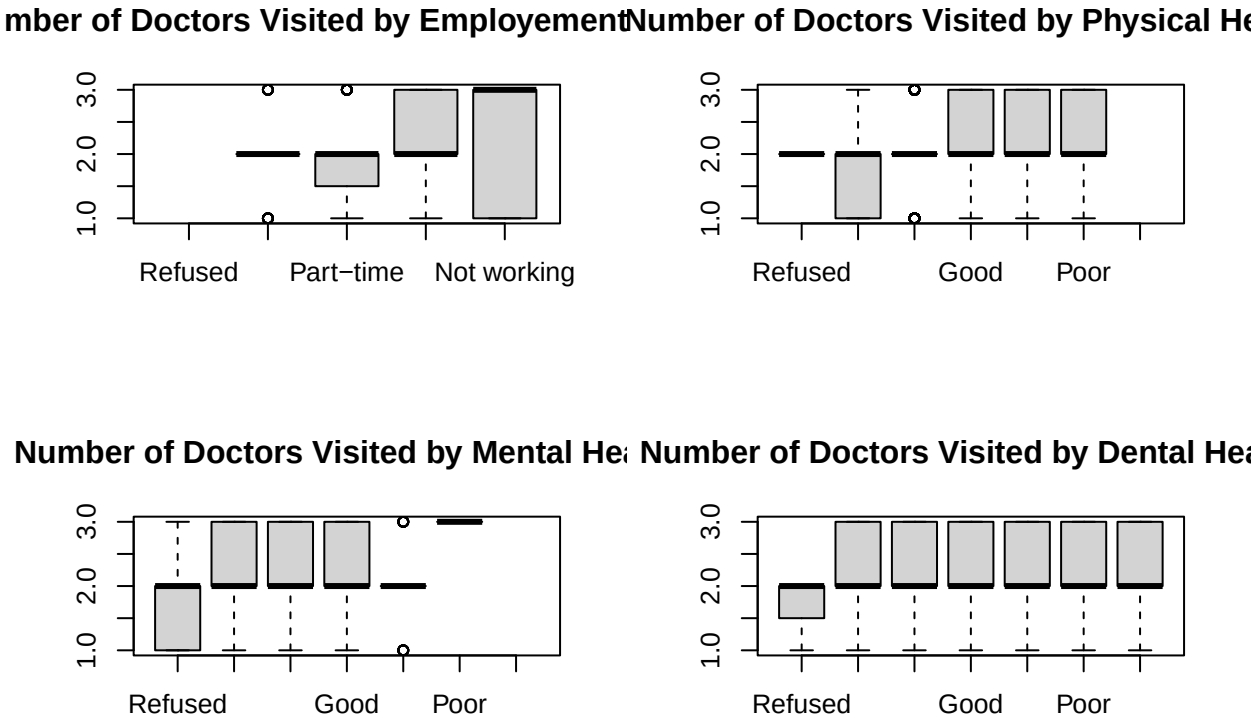
```
plot(npha)
```



Comme attendu, c'est indigeste en vu du nombre de variables présentes dans le jeu de données ainsi que de la présence excessive de variables catégorielles et non numériques.

On va donc essayer de voir les relations entre les variables et le nombre de visites chez le médecin.

```
par(mfrow = c(2, 2))
boxplot(split(npha$Number.of.Doctors.Visited, npha$Employment), main = "Number of Doctors Visited by Emp
boxplot(split(npha$Number.of.Doctors.Visited, npha$Physical.Health), main = "Number of Doctors Visited
boxplot(split(npha$Number.of.Doctors.Visited, npha$Mental.Health), main = "Number of Doctors Visited by
boxplot(split(npha$Number.of.Doctors.Visited, npha$Dental.Health), main = "Number of Doctors Visited by
```



Interprétation des boxplots

Les boxplots révèlent des tendances intéressantes mais ne montrent pas de différences marquées entre les groupes. On observe une légère augmentation du nombre de consultations médicales chez les individus ayant une moins bonne santé physique et mentale, bien que la variabilité reste importante. De même, ceux ayant une mauvaise santé dentaire semblent consulter plus fréquemment, mais l'écart entre les groupes reste modéré. Concernant l'emploi, les personnes retraitées ou sans emploi semblent légèrement plus enclines à consulter un médecin que celles en activité, bien que la différence ne soit pas significative. Globalement, ces distributions suggèrent des tendances faibles mais ne permettent pas d'identifier des facteurs prédictifs forts du nombre de consultations médicales.

3- Classification non supervisée :

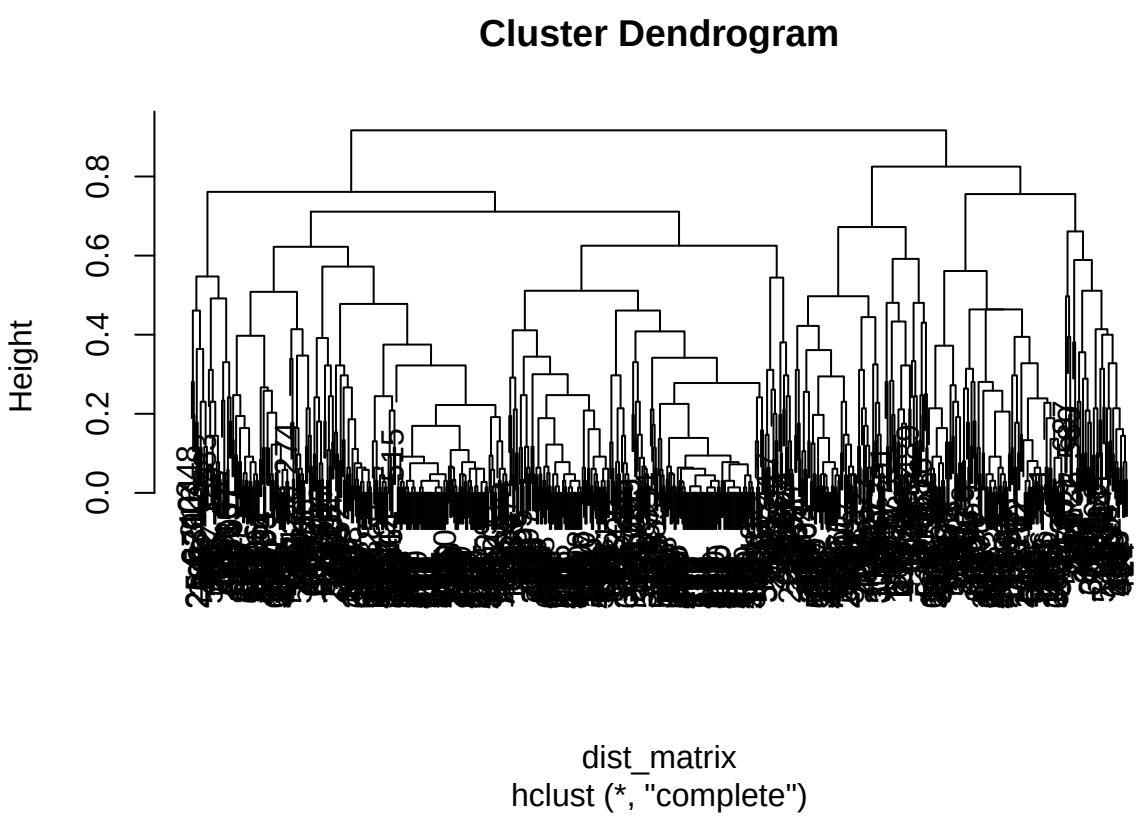
On effectue une analyse de regroupement hiérarchique et un clustering PAM pour segmenter les données sans inclure la variable cible, puis on visualise les résultats de chaque méthode.

```
par(mfrow = c(1, 1))
# Suppression de la variable cible (Number.of.Doctors.Visited)
npha_sans_visites <- npha[, -1]
summary(npha_sans_visites)
```

```
##   Physical.Health   Mental.Health   Dental.Health       Employment
##  Refused  :  1    Refused  : 10   Refused  :  4   Refused    :  0
##  Excellent: 36    Excellent:219   Excellent: 66   Full-time  : 50
```

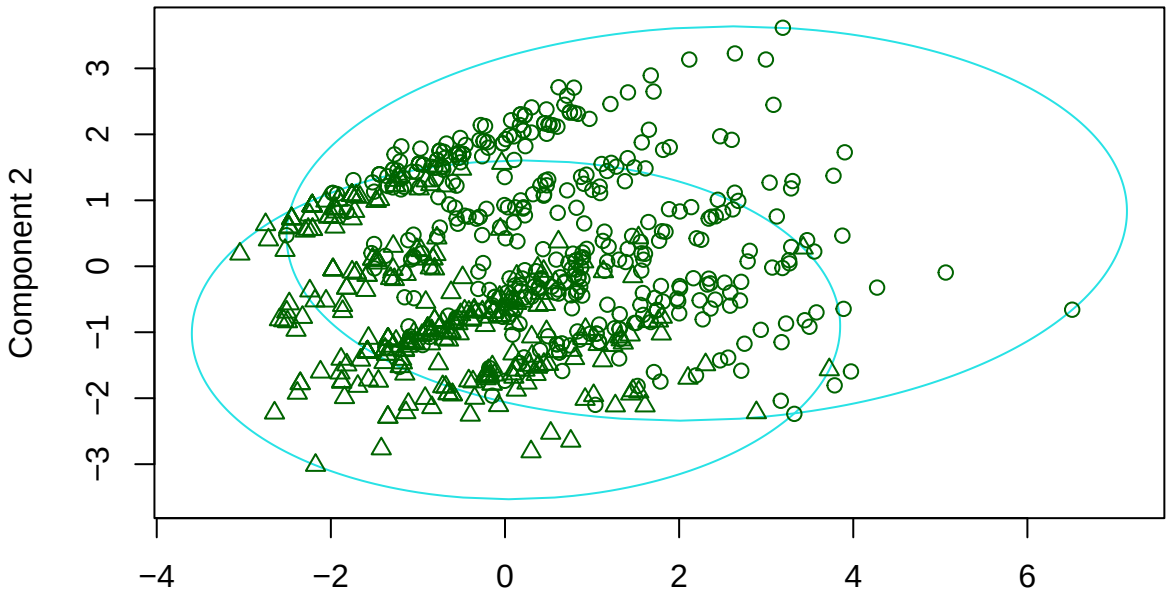
```
##  Very Good:239    Very Good:282   Very Good:215   Part-time  : 55
##  Good     :291    Good     :167   Good     :208   Retired    :592
##  Fair     :126    Fair     : 34   Fair     :127   Not working: 17
##  Poor     : 21    Poor     :  2   Poor     : 39
##  Very Poor:  0    Very Poor:  0   Very Poor: 55
##  Stress.Keeps.Patient.from.Sleeping Medication.Keeps.Patient.from.Sleeping
##  No :537                            No :674
##  Yes:177                            Yes: 40
##
##
##
##
##
##  Pain.Keeps.Patient.from.Sleeping Bathroom.Needs.Keeps.Patient.from.Sleeping
##  No :558                          No :354
##  Yes:156                          Yes:360
##
##
##
##
##
##  Unknown.Keeps.Patient.from.Sleeping Trouble.Sleeping
##  No :416                             Refused:  2
##  Yes:298                             No     : 62
##                                      A bit  :291
##                                      Yes    :359
##
##
##
##   Prescription.Sleep.Medication        Race     Gender
##  Refused         :  3          Not asked:  0   M:321
##  Use regularly   : 38          Refused  :578   F:393
##  Use occasionally: 34          White    : 52
##  Do not use      :639          Black    : 20
##                                Other    : 44
##                                Hispanic : 20
##                                2+ Races :  0
```

```r
dist_matrix <- daisy(npha_sans_visites[, -ncol(npha_sans_visites)])
hclust_result <- hclust(dist_matrix)
plot(hclust_result)
```

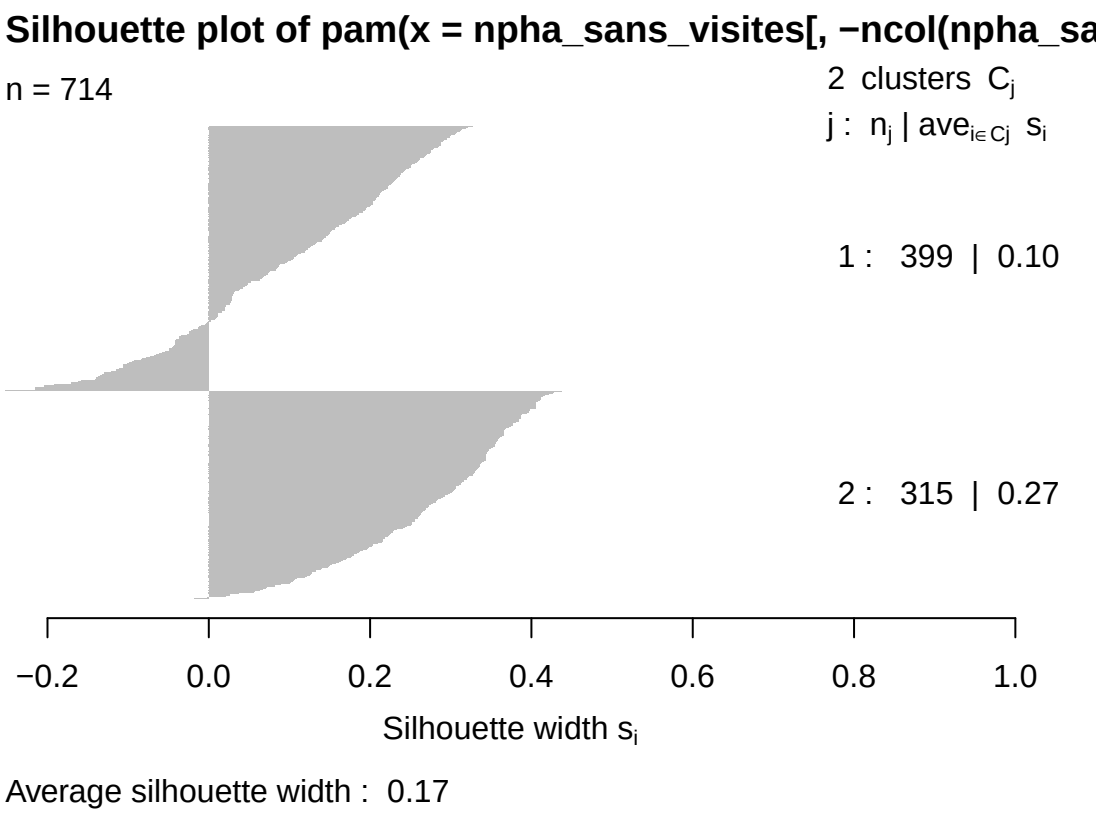



```
pam_result <- pam(npha_sans_visites[, -ncol(npha_sans_visites)], k = 2)
plot(pam_result)
```

clusplot(pam(x = npha_sans_visites[, −ncol(npha_sans_visites)], k =



These two components explain 31.54 % of the point variability.



Interprétation du graphique de clustering PAM

Le graphique de clustering PAM met en évidence deux groupes principaux parmi les observations. Cependant, la séparation entre ces clusters n'est pas nette, indiquant une certaine hétérogénéité au sein des groupes. L'explication de la variance à hauteur de 31,54 % suggère que les deux premières composantes principales ne capturent qu'une partie limitée des informations contenues dans les données. Cette faible variance implique que d'autres dimensions pourraient être nécessaires pour mieux différencier les groupes. De plus, la dispersion des points montre que certains individus sont proches de la frontière entre les clusters, suggérant que les variables choisies ne permettent pas de segmenter clairement la population analysée.

Interprétation du Silhouette Plot

Le Silhouette Plot révèle une cohésion interne relativement faible des clusters, avec une valeur moyenne de 0,17. Ce score indique que de nombreuses observations se situent à la limite de leur groupe, ce qui traduit une séparation imparfaite entre les clusters. En particulier, le premier cluster présente une silhouette moyenne plus basse, ce qui signifie que ses individus sont plus dispersés et donc moins homogènes. À l'inverse, le second cluster semble mieux défini, bien que sa cohésion reste modérée. Globalement, ces résultats suggèrent que le choix du nombre de clusters pourrait être optimisé ou que certaines variables devraient être réévaluées pour améliorer la qualité de la classification.

4- Classification supervisée:

On commence par récuperer les indices pour chaque catégorie de la variable cible. On a observé que bien qu'il n'y ait aucune classe extrémement rare, leur distribution n'est pas équilibrée. On essaie donc dans

un premier temps avec un échantillon de 50 observations par classe pour voir si le modèle arrive à prédire correctement les classes.

```
# Ici aucun class est assez rare pour faire un max d'une classe

indices_1 <- which(npha$Number.of.Doctors.Visited == "0-1")
indices_2 <- which(npha$Number.of.Doctors.Visited == "2-3")
indices_3 <- which(npha$Number.of.Doctors.Visited == "4+")

sample_1 <- sample(indices_1, 50)
sample_2 <- sample(indices_2, 50)
sample_3 <- sample(indices_3, 50)
sub <- c(sample_1, sample_2, sample_3)
```

On vérifie ici qu'une des classes n'est pas trop sous-représentée.

```
length(indices_1)
```

```
## [1] 131
```

```
length(indices_2)
```

```
## [1] 372
```

```
length(indices_3)
```

```
## [1] 211
```

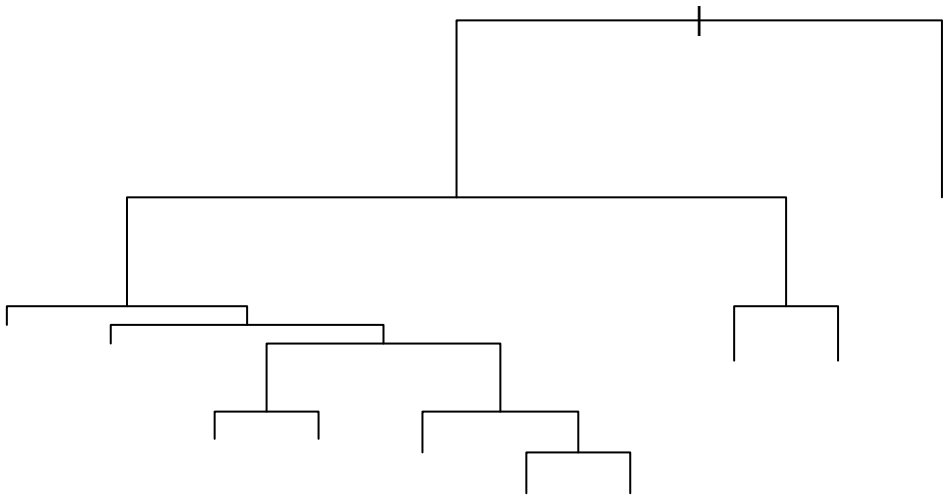
On entraîne un modèle de classification rpart sur un sous-ensemble de données et on évalue ses prédictions en comparant les résultats sur un jeu de test excluant ces mêmes indices.

```
fit <- rpart(npha$Number.of.Doctors.Visited ~ ., data = npha, subset = sub)
fit
```

```
## n= 150
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 150 100 0-1 (0.33333333 0.33333333 0.33333333)
##     2) Prescription.Sleep.Medication=Refused,Do not use 129  80 0-1 (0.37984496 0.34108527 0.2790697
##       4) Physical.Health=Refused,Excellent,Very Good,Good 99  55 0-1 (0.44444444 0.31313131 0.2424241
##         8) Mental.Health=Good,Fair,Poor,Very Poor 10   2 0-1 (0.80000000 0.10000000 0.10000000) *
##         9) Mental.Health=Refused,Excellent,Very Good 89  53 0-1 (0.40449438 0.33707865 0.25842697)
##          18) Physical.Health=Refused,Excellent 7   1 0-1 (0.85714286 0.14285714 0.00000000) *
##          19) Physical.Health=Very Good,Good,Fair,Poor,Very Poor 82  52 0-1 (0.36585366 0.35365854 0.1
##            38) Unknown.Keeps.Patient.from.Sleeping=Yes 40  21 0-1 (0.47500000 0.32500000 0.20000000)
##              76) Physical.Health=Good,Fair,Poor,Very Poor 16   6 0-1 (0.62500000 0.12500000 0.2500000
##              77) Physical.Health=Refused,Excellent,Very Good 24  13 2-3 (0.37500000 0.45833333 0.166
##            39) Unknown.Keeps.Patient.from.Sleeping=No 42  26 2-3 (0.26190476 0.38095238 0.35714286)
##              78) Stress.Keeps.Patient.from.Sleeping=Yes 18  10 0-1 (0.44444444 0.33333333 0.22222222)
```

```
##              79) Stress.Keeps.Patient.from.Sleeping=No 24  13 4+ (0.12500000 0.41666667 0.45833333)
##               158) Trouble.Sleeping=Yes 15   7 2-3 (0.13333333 0.53333333 0.33333333) *
##               159) Trouble.Sleeping=A bit 9   3 4+ (0.11111111 0.22222222 0.66666667) *
##       5) Physical.Health=Fair,Poor,Very Poor 30  17 2-3 (0.16666667 0.43333333 0.40000000)
##        10) Dental.Health=Fair,Poor,Very Poor 20  10 2-3 (0.25000000 0.50000000 0.25000000) *
##        11) Dental.Health=Refused,Excellent,Very Good,Good 10   3 4+ (0.00000000 0.30000000 0.7000000
##     3) Prescription.Sleep.Medication=Use regularly,Use occasionally 21   7 4+ (0.04761905 0.28571429
```

```
plot(fit)
```



```
res <- table(predict(fit, npha[-sub, ], type = "class"), npha[-sub, "Number.of.Doctors.Visited"])
res
```

```
##
##       0-1 2-3  4+
##   0-1  31 136  69
##   2-3  34 119  52
##   4+   16  67  40
```

On calcule le taux d'erreur du modèle en comparant les prédictions correctes aux résultats totaux, puis on l'affiche sous forme de pourcentage.

```
err <- (1 - sum(diag(res)) / sum(res)) * 100
cat("Le taux d'erreur est de :", err, "%\n")
```

```
## Le taux d'erreur est de : 66.31206 %
```

Impact de l'équilibre des classes sur la performance du modèle

Au départ, l'entraînement du modèle a été réalisé en sélectionnant un nombre fixe d'observations par classe afin de garantir une répartition équilibrée entre les catégories. Cette approche permet d'éviter qu'une classe majoritaire domine l'apprentissage, ce qui pourrait biaiser les prédictions et réduire la capacité du modèle à identifier correctement les classes moins représentées.

Cependant, cette méthode ne reflète pas la distribution réelle des données. C'est pourquoi nous avons testé une seconde approche où chaque classe est échantillonnée selon un même pourcentage de ses observations totales. Cette méthode permet au modèle d'apprendre à partir d'une répartition plus représentative de la réalité, ce qui peut améliorer sa capacité de généralisation. Comparer ces deux approches permet d'évaluer si un équilibre artificiel améliore la précision ou si une répartition proportionnelle aux données initiales est plus pertinente.

```
indices_1 <- which(npha$Number.of.Doctors.Visited == "0-1")
indices_2 <- which(npha$Number.of.Doctors.Visited == "2-3")
indices_3 <- which(npha$Number.of.Doctors.Visited == "4+")

sample_1 <- sample(indices_1, round(0.6 * length(indices_1), digits = 0))
sample_2 <- sample(indices_2, round(0.6 * length(indices_2), digits = 0))
sample_3 <- sample(indices_3, round(0.6 * length(indices_3), digits = 0))
sub <- c(sample_1, sample_2, sample_3)

fit <- rpart(npha$Number.of.Doctors.Visited ~ ., data = npha, subset = sub)
res <- table(predict(fit, npha[-sub, ], type = "class"), npha[-sub, "Number.of.Doctors.Visited"])

err <- (1 - sum(diag(res)) / sum(res)) * 100
cat("Le taux d'erreur est de :", err, "%\n")
```

```
## Le taux d'erreur est de : 51.57895 %
```

Les résultats montrent que l'entraînement avec un échantillonnage proportionnel réduit significativement le taux d'erreur par rapport à un échantillonnage fixe. Toutefois, cela ne confirme pas que respecter la distribution naturelle des données permet au modèle de mieux généraliser car étant donné que le taux d'erreur reste très élevé, il est très probable que cette baisse soit en fait due à une sur-représentation des classes majoritaires dans l'échantillon d'entraînement, ce qui entraine une baisse artificielle du taux d'erreur en biaisant le modèle.

Conclusion:

Les résultats de l'étude montrent que les variables analysées nee semblent pas expliquer le nombre de consultations médicales des personnes âgées. Les tentatives de classification ont révélé une segmentation peu claire des groupes, avec une faible cohésion interne et un fort chevauchement entre les observations.

L'analyse des clusters a indiqué une variance expliquée limitée (31,54 %) et un score de silhouette faible (0,17), suggérant une séparation imparfaite des groupes. Les individus ne se distinguent pas nettement en fonction des variables étudiées, ce qui remet en question leur pertinence pour la prédiction du comportement médical.

De plus, le modèle supervisé testé a affiché un taux d'erreur élevé (66.13475 % / 51.92982 %), confirmant la difficulté à établir une relation fiable entre les caractéristiques des patients et leur fréquence de consultation. Aucune variable ne semble exercer une influence prédictive suffisante pour permettre une classification efficace. Additionnellement, le fait qu'un données d'entrainement propotionnelles, le taux d'erreur soit proche

de 2/3 semble souligner d'autant plus une totale absence de lien, car c'est ce qu'on pourrait esperer en cas de prédiction strictement aléatoire.

En conclusion, les résultats indiquent que les données disponibles ne sont pas assez discriminantes pour prédire avec précision le nombre de visites médicales. Des variables supplémentaires ou une approche différente seraient nécessaires pour améliorer la capacité prédictive des modèles.