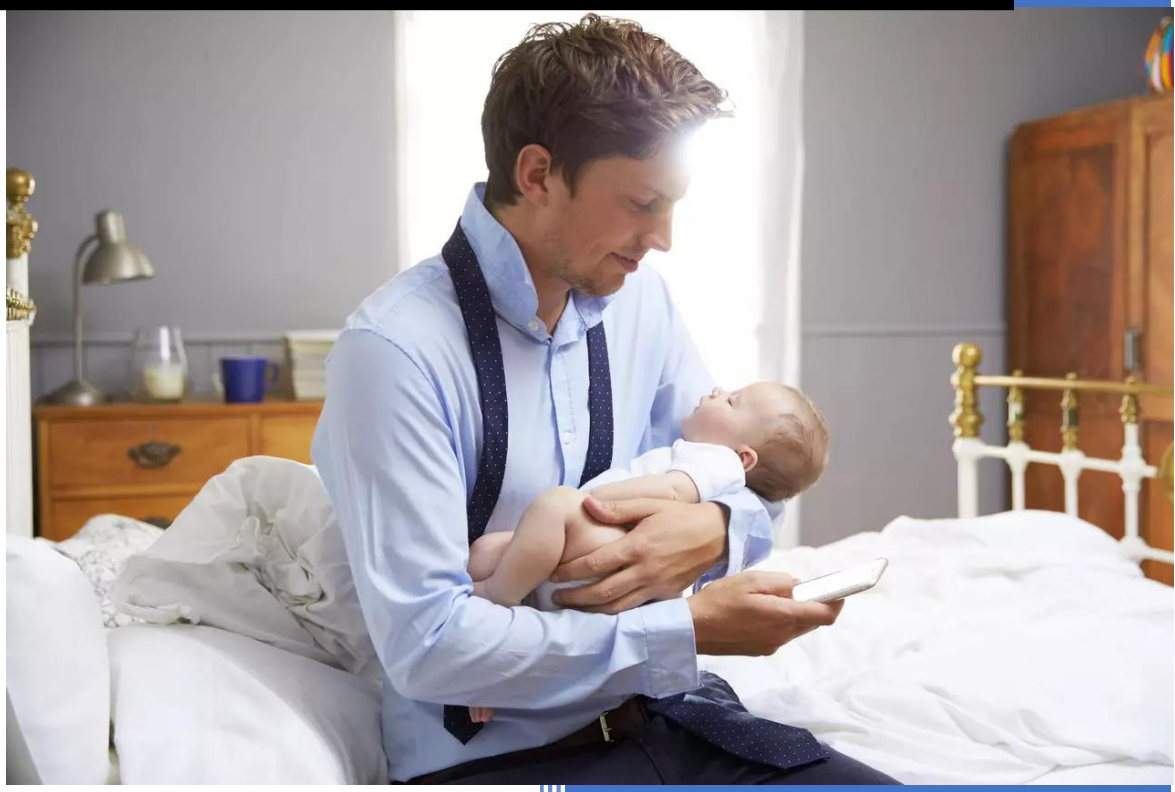
Rapport d’analyse décisionnelle :Fertilité



GRASINA Willy

TIYADJOWE Antoine

Table des matières

1. Motivation et positionnement du projet :

Problématique business : L'infertilité masculine est en hausse. Des causes épigénétiques semblent être la raison.[1]

Traduction de la problématique en objectif data : Besoin de prédire la fertilité (ou l'infertilité) des hommes à partir des données provenant de leur mode de vie et de leur environnement social.

Intérêt : Cette approche apporte un complément aux tests génétiques d'infertilité. Elle permet de réaliser des tests simples et confortables pour les individus et d'envisager une prévention contre les causes identifiées (si possible).

Problématique data : Détecter les variables expliquant le statut fertile / infertile des individus.

Objectif Data science : *Le but de ce travail est donc de proposer aux experts un modèle permettant de prédire le statut fertile / infertile sur la base de variables non génétiques.*

2. Analyse descriptive :

2.1. Concept de l'analyse descriptive :

L'objectif d'une analyse descriptive est de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses. Pour ce faire, il existe plusieurs méthodes. On peut notamment avoir recours à des histogrammes, des nuages de points ou des diagrammes en boîtes.

2.2. Histogrammes :

Un histogramme est un outil permettant de représenter les données et de les décrire en utilisant des lois statistiques. Grâce à un histogramme, il est ainsi possible d'observer la répartition statistique d'une variable, c'est-à-dire de savoir à quelle fréquence cette variable prend telle ou telle valeur. Evaluer la répartition d'une variable permet d'émettre des hypothèses quant à son rôle dans l'étude d'un phénomène. En effet, si pour un événement donné B, une variable A est répartie de façon équiprobable, alors on peut penser que A et B sont indépendants.

Dans le cas de notre étude (**Figure 1**), on note que la répartition est quasiment équiprobable pour les variables « **Accident or serious Trauma** » et « **Surgical Intervention** ». Cela laisse penser que ces variables ne seront pas des facteurs décisifs dans le résultat du diagnostic. A l'opposé, on note une diminution de la fréquence pour la variable « **Age** » lorsqu'on passe de 0.5 à 1, ce qui permet de d'envisager une potentielle corrélation entre la variable « **Age** » et le résultat du diagnostic.

[1] Auger, Jacques, et al. « Decline in Semen Quality among Fertile Men in Paris during the Past 20 Years ». New England Journal of Medicine, vol. 332, no 5, février 1995, p. 281-85. DOI.org (Crossref), doi:10.1056/NEJM199502023320

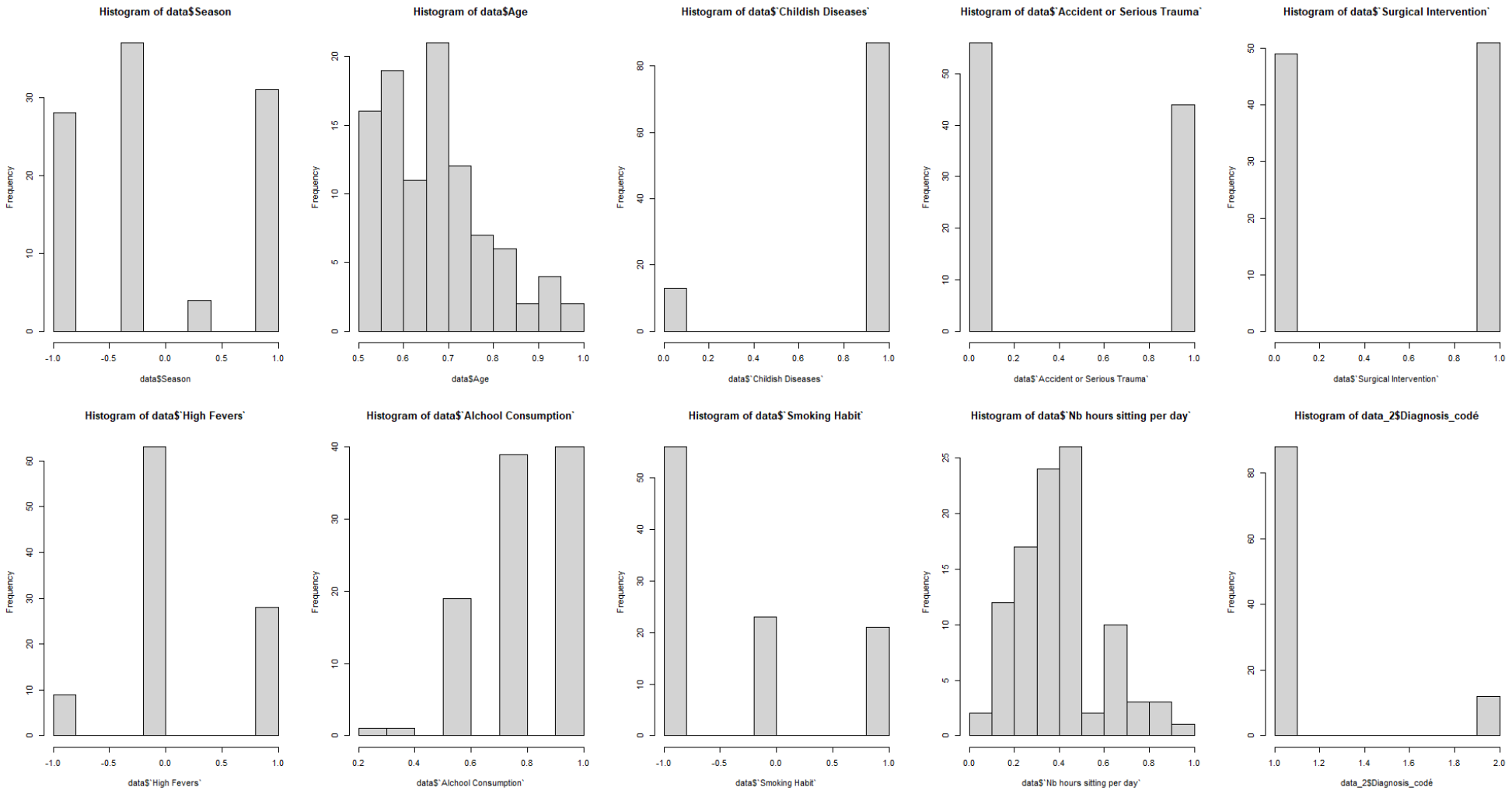


Figure 1 : Histogrammes des variables

2.3. Matrices de nuages de points:

Figure 2: Matrice de nuages de points

La **Figure 2** présente la matrice des nuages de points de notre étude. Elle permet d'identifier les éventuelles corrélations entre les différentes variables.

Pour ce qui est de la corrélation avec la variable à prédire (**Diagnosis**), nous nous sommes intéressés à regarder pour quelles variables, la distribution de « **Diagnosis** » (1= normal, 2 = altered) n'est pas symétrique. A cet effet, on regarde pour la colonne « **Diagnosis** », les variables pour lesquelles la densité des points est équivalente dans chaque strate. Les strates correspondent aux différentes valeurs possibles prises par l'autre variable. On remarque ainsi que les variables « **Age** », « **Alcohol** » et « **Number hours sitting per day** » sont les plus corrélées à la variable « **Diagnosis** ».

On note également des corrélations entre certaines variables qui peuvent être porteuse d'intérêt. Il s'agit plus précisément de la corrélation linéaire négative observée entre les variables « **Age** » et « **Number hours sitting per day** » avec un taux de corrélation de -44% (**Figures 2 & 3**). Cette observation semble signifier que les personnes plus âgées sont celles qui restent le moins assises ce qui peut *a priori* paraître comme contre-intuitif. Cependant, cela peut s'expliquer par le fait que la tranche d'âge prise en compte dans le cadre de cette étude est celle de 18 à 36 ans. Ainsi, à 36 ans on est plutôt très actif sur le plan socio-professionnel (et donc très mobil) alors que les personnes âgées de 18 ans qui sont en majorité des étudiants passent généralement de longues heures assis (pour suivre des cours par exemple).

```
> cor(data_3)
                                  Season         Age Childish Diseases Accident or Serious Trauma Surgical Intervention   High Fevers Alchool Consumption
Season                       1.000000000  0.06540951       -0.17650891                -0.09627403          -0.006210151 -0.2218184462       -0.0412904406
Age                          0.065409511  1.00000000        0.08055115                 0.21595797           0.271944809  0.1202843352       -0.2479401810
Childish Diseases           -0.176508907  0.08055115        1.00000000                 0.16293607          -0.140972342  0.0756445812        0.0385378602
Accident or Serious Trauma  -0.096274033  0.21595797        0.16293607                 1.00000000           0.103165975 -0.0822779461       -0.2427216440
Surgical Intervention       -0.006210151  0.27194481       -0.14097234                 0.10316597           1.000000000 -0.2315979869       -0.0758576000
High Fevers                 -0.221818446  0.12028434        0.07564458                -0.08227795          -0.231597987  1.0000000000       -0.0008307052
Alchool Consumption         -0.041290441 -0.24794018        0.03853786                -0.24272164          -0.075857600 -0.0008307052        1.0000000000
Smoking Habit               -0.028084754  0.07258127        0.09053466                 0.11015709          -0.053448491 -0.0075273125       -0.1849255601
Nb hours sitting per day    -0.019020983 -0.44245195       -0.14776062                 0.01312182          -0.192726024 -0.1510914297        0.1113714891
Diagnosis_codé               0.192416862  0.11522900       -0.04026145                -0.14134568           0.054171091 -0.1214212441       -0.1447602258
                            Smoking Habit Nb hours sitting per day Diagnosis_codé
Season                       -0.028084754            -0.019020983     0.19241686
Age                           0.072581268            -0.44245195      0.11522900
Childish Diseases             0.090534664            -0.14776062     -0.04026145
Accident or Serious Trauma    0.110157086             0.01312182     -0.14134568
Surgical Intervention        -0.053448491            -0.19272602      0.05417109
High Fevers                  -0.007527313            -0.15109143     -0.12142124
Alchool Consumption          -0.184925560             0.11137149     -0.14476023
Smoking Habit                 1.000000000            -0.10600690      0.04589120
Nb hours sitting per day     -0.106006901             1.00000000      0.02296422
Diagnosis_codé                0.045891199             0.02296422      1.00000000
> |
```

Figure 3 : Rapport des taux corrélations entre variables

2.4. Boxplots :

Encore appelés « boîtes à moustaches », les boxplots servent à visualiser les positions relatives de la médiane ainsi que celles du premier et du troisième quartile d'une distribution.

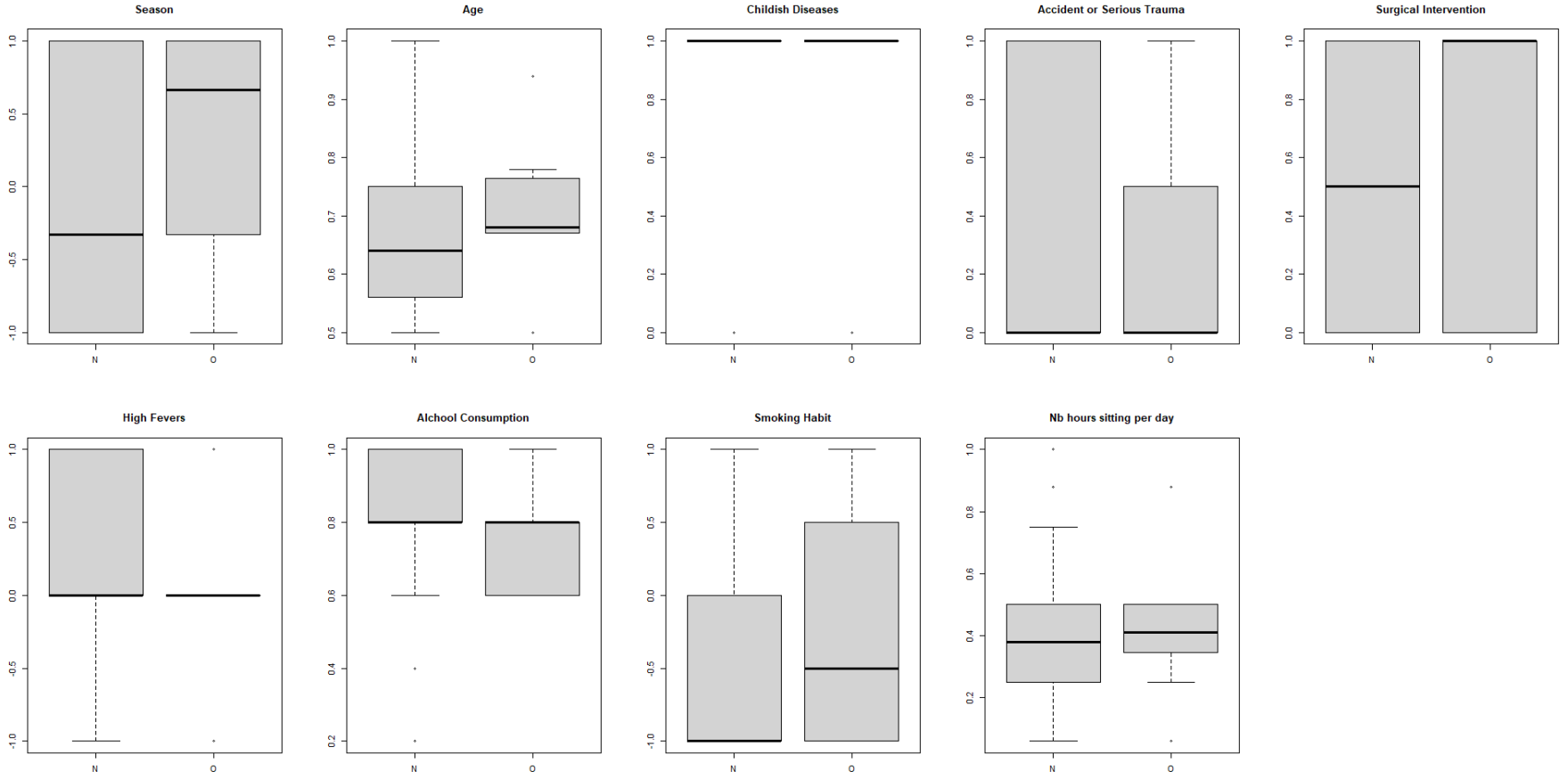


Figure 4 : Diagrammes en boîtes

De ces boxplots, on peut tirer les conclusions suivantes :

Age : On peut observer que les individus ayant un diagnostic "O", sont en moyenne plus âgés que les individus sains.

Alcohol : Les individus sains consomment en moyenne moins d'alcool (1,0.8 correspondant à peu d'alcool consommé) que les individus atteints.

Number hours sitting per day : Il semblerait que les personnes restant assises plus longtemps souffriraient plus souvent d'infertilité.

Season : Durant les saisons les plus chaudes (Automne, Eté,)= saisons plus chaudes, il est logique de s'attendre à une baisse de la fertilité, étant donné qu'une température trop élevée peut détruire les spermatozoïdes.

Accidents and serious trauma : On constate qu'il y a moins d'accidents chez les personnes saines

Smoking habit : Les individus fument en moyenne moins que les individus atteints.

3. Classification non supervisée :

3.1. K-means :

3.1.1. K-means préliminaire :

Nous avons tenté un Kmeans avec 2 classes étant donné que la variable "Diagnostic" possède 2 modalités. Le but était de voir si le K-means pouvait distinguer les 2 modalités N/O du diagnostic. Cependant, le résultat était décevant. En effet, l'inertie était de 24,4 %. Le K Means ne semble ainsi pas pouvoir distinguer les 2 groupes N/O.

Quand on visualise les clusters, on voit en effet, que les données semblent être très compacte. Il ne semble pas avoir de groupes facilement identifiables.

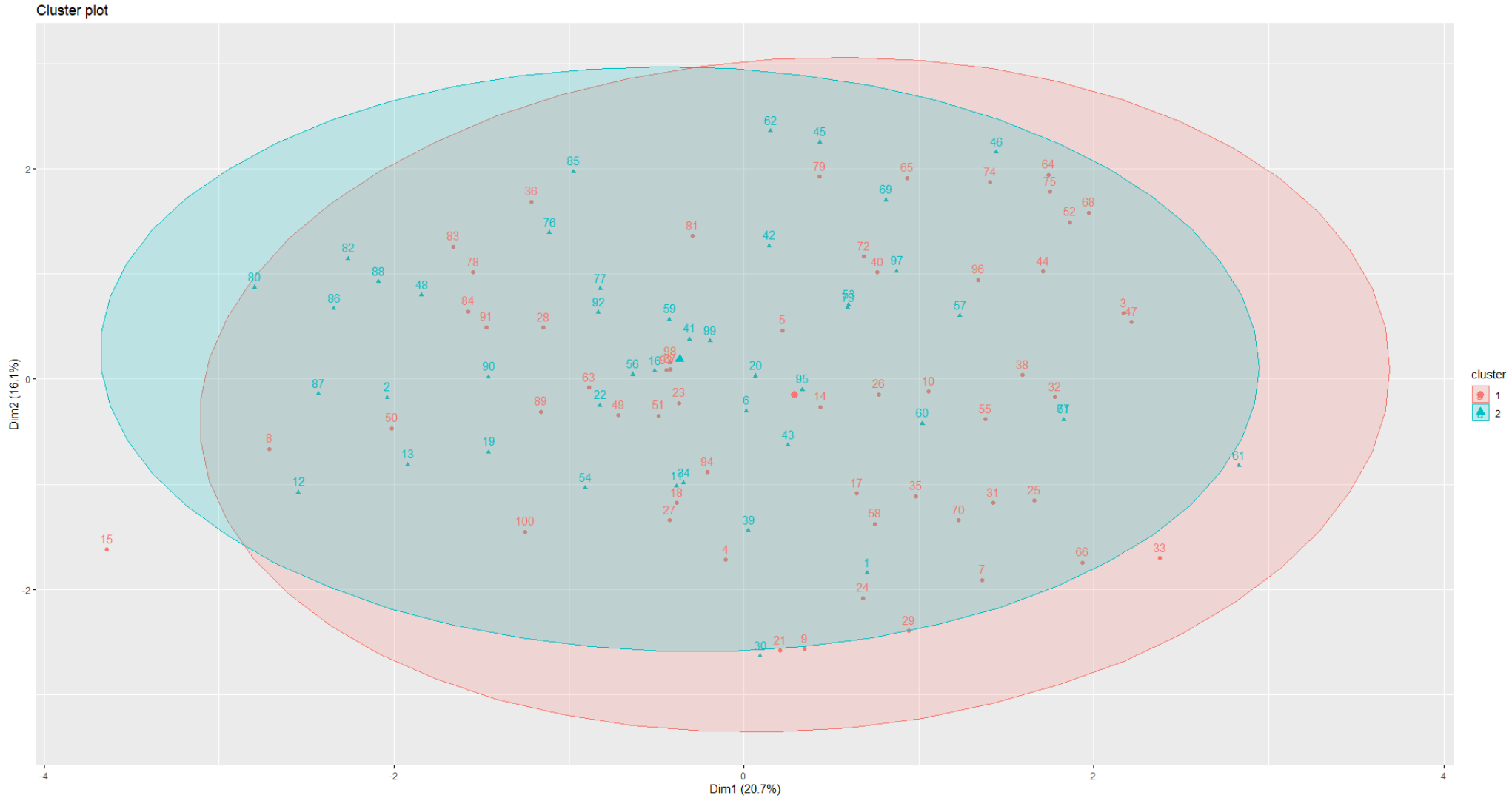


Figure 5 : Visualisation K means à 2 classes

3.1.2. Recherche du bon nombre de classes :

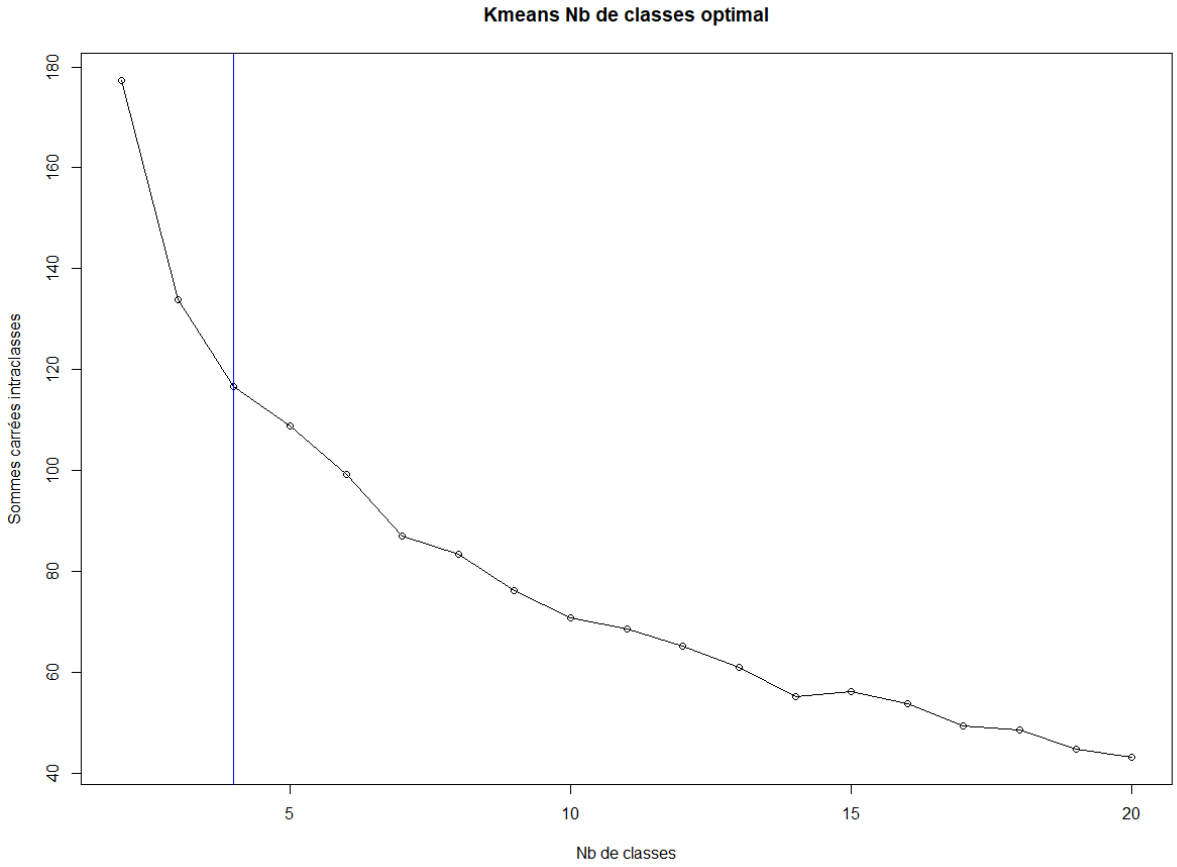


Figure 6 : Détermination du nombre de classes

Pour déterminer le nombre de classes optimal dans un K-means on regarde l'évolution de la somme des carrés intra classe en fonction du nombre de classes. En effet, cette dernière mesure la distance moyenne entre les individus d'une même classe. Moins elle est élevée, plus le clustering est de bonne qualité.

Ici, on voit que la somme des carrés diminue rapidement puis Semble atteindre un coude entre 4 et 6 classes. Etant donné la structure fortement compacte des données, nous avons jugé inutile de séparer les individus en un nombre important de classes. Nous avons donc retenu 4 classes pour le K-means.

3.1.3. Visualisation du K-means final à 4 classes :

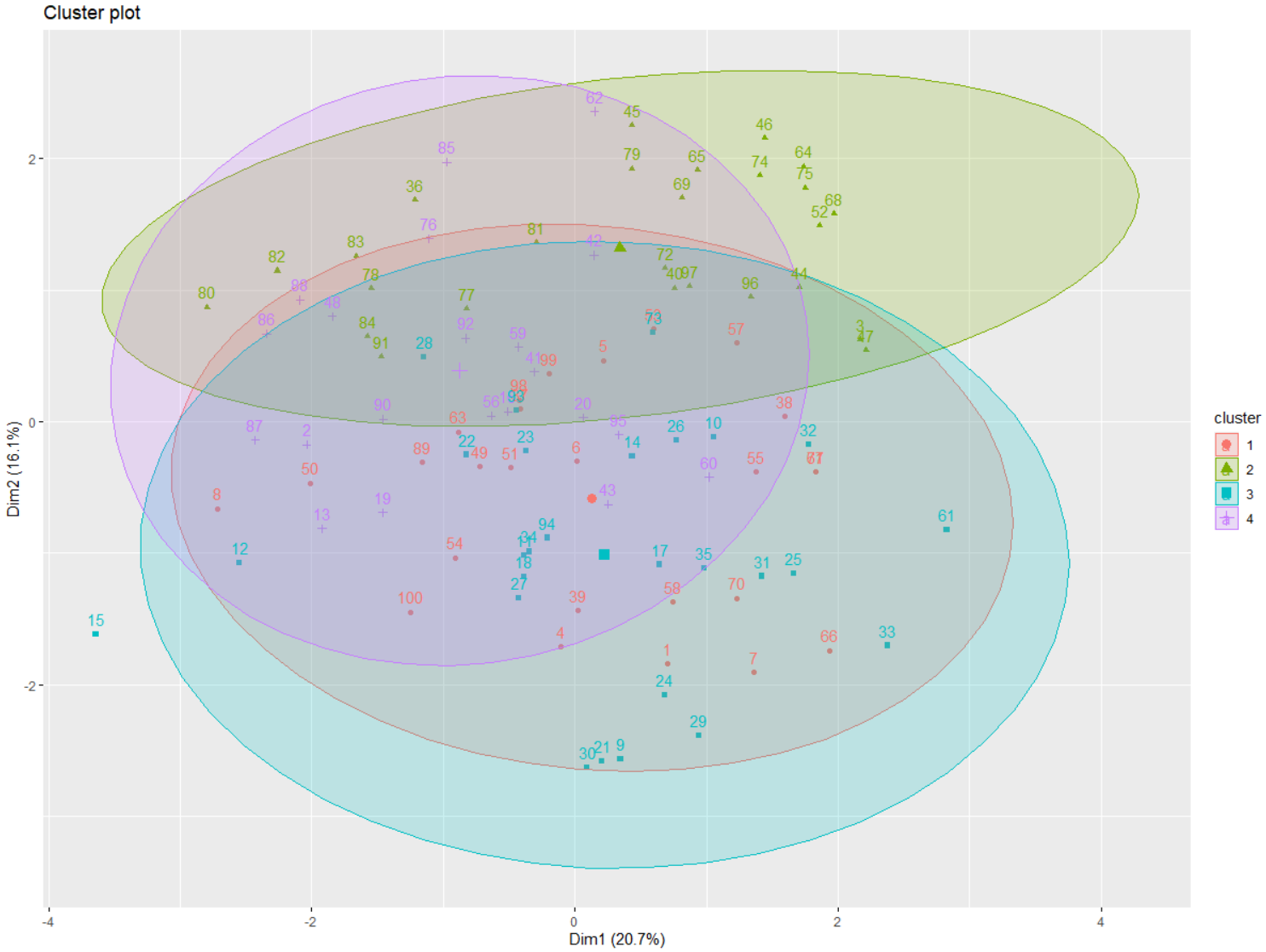


Figure 7 : Visualisation K means à 4 classes

3.1.4. Extraction des profils types :

```
> K_means_final
K-means clustering with 4 clusters of sizes 26, 26, 27, 21

Cluster means:
       Season       Age Childish Diseases Accident or Serious Trauma Surgical Intervention High Fevers Alchool Consumption Smoking Habit
1 -0.51038462 0.6523077         0.7692308                 0.6153846            0.7692308 -0.23076923           0.8307692    -0.6538462
2 -0.71692308 0.6573077         1.0000000                 0.3461538            0.2307692  0.76923077           0.8692308    -0.7307692
3  0.95037037 0.6688889         0.7777778                 0.2962963            0.5555556  0.03703704           0.8148148    -0.7407407
4 -0.07809524 0.7042857         0.9523810                 0.5238095            0.4761905  0.19047619           0.8095238     1.0000000
  Nb hours sitting per day
1                0.4680769
2                0.4134615
3                0.4003704
4                0.3309524
```

Figure 8 : Centroïdes du K-means

Classe 1	Hiver-printemps	30 ans	N'ont en général pas eu de maladies infantiles	N'ont en général pas eu de traumatismes	N'ont en général pas subi d'opérations	Ont en général eu des fièvres il y a plus de 3 mois	Boivent en général 1 fois par semaine	Ne fument en moyenne pas	Restent assis 7 heures par jour
Classe 2	Hiver-printemps	30 ans	N'ont pas eu de malades infantiles	Ont eu général eu des traumatismes	Ont eu général subi des opérations	N'ont en général pas eu de fièvres	Boivent en général 1 fois par semaine	Ne fument en moyenne pas	Restent assis 6,5 heures par jour
Classe 3	Automne	30 ans	N'ont en général pas eu de maladies infantiles	Ont eu général eu des traumatismes	Egalité	Ont en général eu des fièvres il y a plus de 3 mois	Boivent en général 1 fois par semaine	Ne fument en moyenne pas	Restent assis 6,5 heures par jour
Classe 4	Printemps-été	31 ans	N'ont pas eu de malades infantiles	Egalité	Egalité	Ont en général eu des fièvres il y a plus de 3 mois	Boivent en général 1 fois par semaine	Fument tous les jours	Restent assis 5 heures par jour

3.2. PAM :

3.2.1. Détermination du bon nombre de classes :

Pour déterminer le nombre de classes optimal dans un PAM, on regarde l'évolution de la valeur silhouette globale en fonction du nombre de classes. En effet, cette dernière mesure, plus ou moins, le pourcentage d'individus bien classés. Ici, on voit que la valeur silhouette augmente rapidement puis diminue à partir de 5 classes. Cette chute continue jusqu'à 8 classes puis remontre à partir de là. Etant donné que nous avions obtenu un nombre optimal de classes de 5 dans le K-means, il semble logique de prendre un nombre proche pour le PAM. Au vu de la valeur silhouette, nous avons choisi de prendre 5 classes.

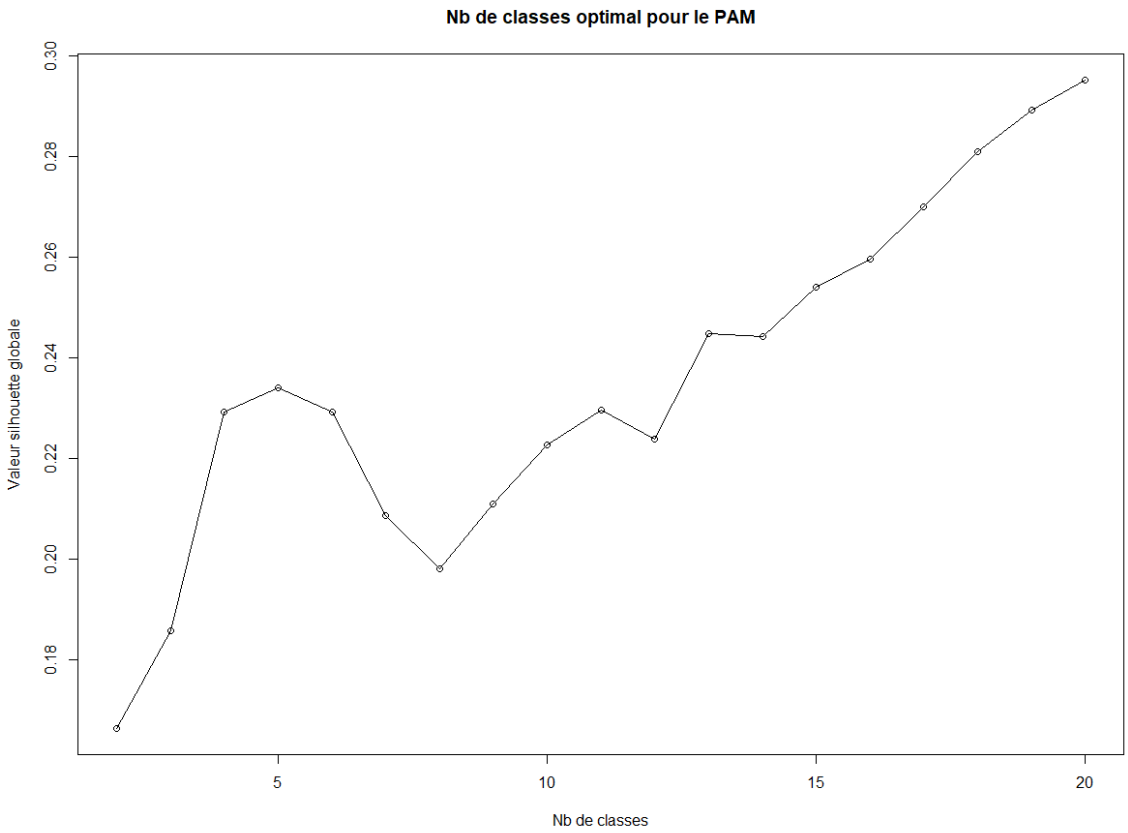


Figure 9 : PAM, détermination du nombre de classes

Malheureusement, de même que le Kmeans, les résultats sont assez décevants. On obtient, en effet, une valeur silhouette de globale de 0.23. Cette dernière oscillant entre 0.5 pour les individus les mieux classés et des valeurs négatives pour les pires résultats comme en témoigne le graphe silhouette suivant.

Les résultats restent peu convaincants. Il semblerait que les données n'aient qu'une structure artificielle au mieux.

3.2.2. Graphe silhouette :

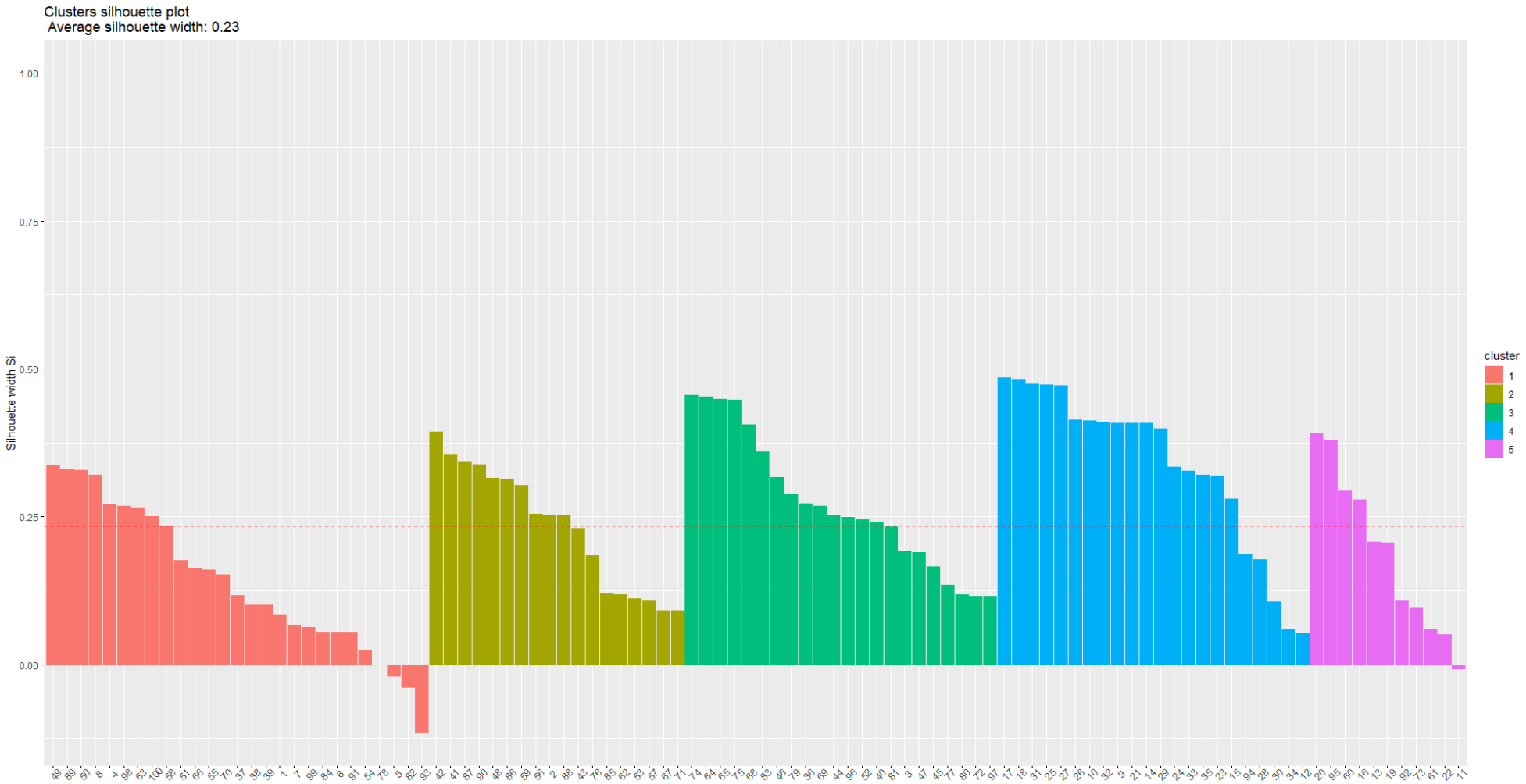


Figure 10 : PAM à 5 classes, graphe silhouette

3.2.3. Visualisation du PAM à 5 classes :

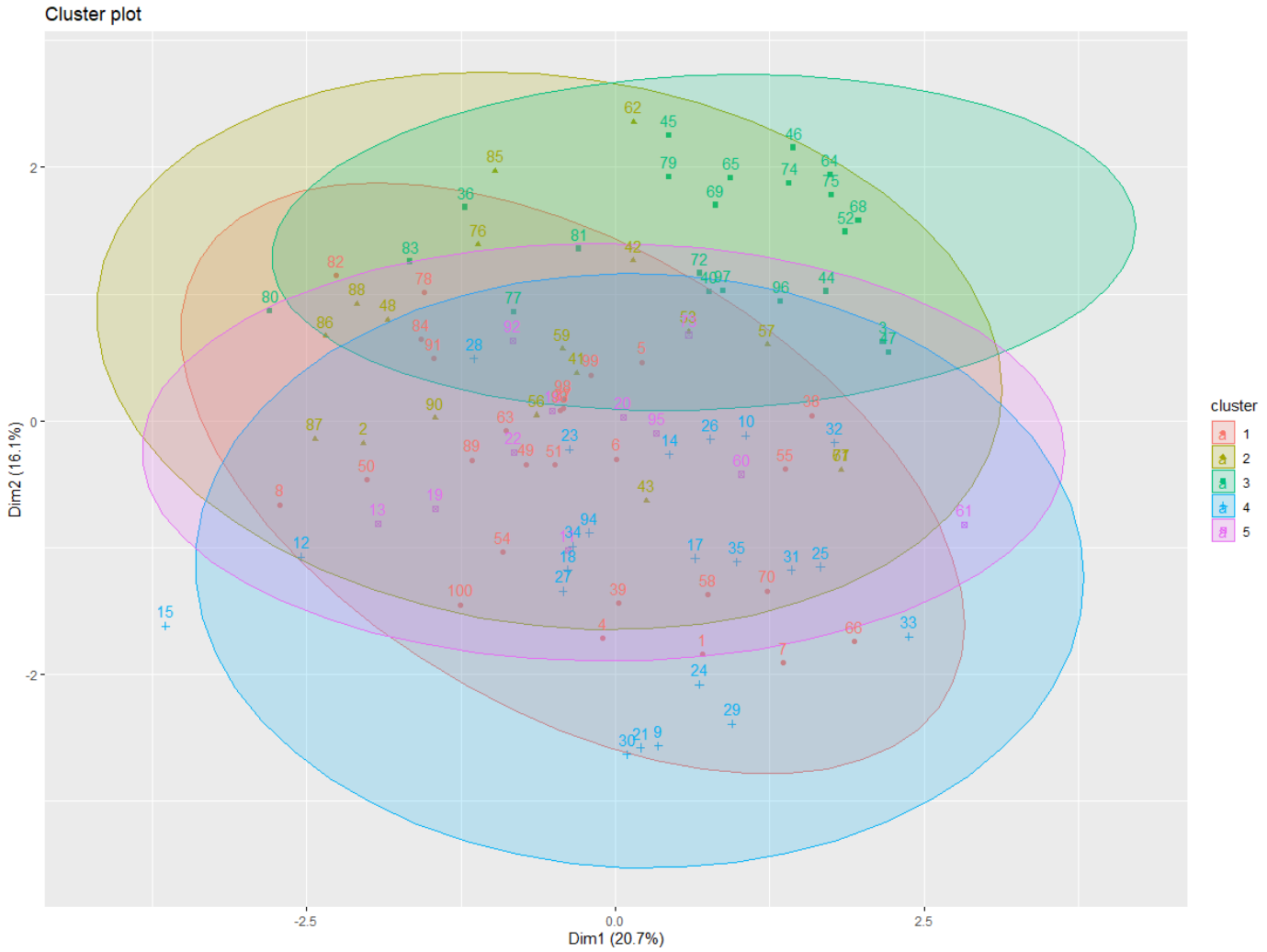


Figure 11 : Visualisation PAM à 5 classes

3.2.4. Extraction des profils types :

```
> PAM_final$medoids
     Season  Age Childish Diseases Accident or Serious Trauma Surgical Intervention High Fevers Alchool Consumption Smoking Habit
[1,]  -0.33 0.64                 1                            1                    1           0                0.8            -1
[2,]  -0.33 0.72                 1                            1                    0           0                0.6             1
[3,]  -1.00 0.50                 1                            0                    0           1                0.8            -1
[4,]   1.00 0.64                 1                            0                    1           0                1.0            -1
[5,]   1.00 0.67                 1                            0                    0           0                0.8             1
     Nb hours sitting per day
[1,]                     0.31
[2,]                     0.19
[3,]                     0.44
[4,]                     0.38
[5,]                     0.38
>
```

Figure 12: Médoïdes du PAM

Classe 1	Printemps	29 ans	N'a pas eu de maladie infantiles	N'a pas eu de traumatismes	N'a pas subi d'interventions	A eu des fièvres dans il y a moins de 3 mois	Boit 1 fois par semaine	Ne fume pas	Reste assit 5 heures par jour
Classe 2	Printemps	31 ans	N'a pas eu de maladie infantiles	N'a pas eu de traumatismes	A subit des interventions	A eu des fièvres dans il y a moins de 3 mois	Boit plusieurs fois par semaine	Fume tous les jours	Reste assit 3 heures par jour
Classe 3	Hiver	27 ans	N'a pas eu de maladie infantiles	A eu des traumatismes	A subit des interventions	A eu des fièvres dans il y a plus de 3 mois	Boit 1 fois par semaine	Ne fume pas	Reste assit 7 heures par jour
Classe 4	Automne	29 ans	N'a pas eu de maladie infantiles	A eu des traumatismes	N'a pas subi d'interventions	A eu des fièvres dans il y a moins de 3 mois	Ne boit pas	Ne fume pas	Reste assit 6 heures par jour
Classe 5	Automne	30 ans	N'a pas eu de maladie infantiles	A eu des traumatismes	A subit des interventions	A eu des fièvres dans il y a moins de 3 mois	Boit 1 fois par semaine	Fume tous les jours	Reste assit 6 heures par jour

On remarque tout de suite que les maladies infantiles ne sont pas une variable caractéristique pour le PAM étant donné que les médoïdes de chaque classe présentent la même modalité.

-

3.3. CAH :

3.3.1. Principe de la CAH:

La classification ascendante hiérarchique est une méthode de clustering permettant de regrouper les individus selon leur dissimilarité. Initialement, chaque individu forme une classe, soit n classes. À chaque étape, on fusionne deux classes, réduisant ainsi le nombre de classes. Les deux classes choisies pour être fusionnées sont celles qui sont les plus « proches », en d'autres termes, celles dont la dissimilarité entre elles est minimale, cette valeur de dissimilarité est appelée indice d'agrégation. Une fois le clustering terminé, il est possible de construire un dendrogramme représentant les différents individus ainsi que leur distance d'agrégation.

3.3.2. Dendrogramme :

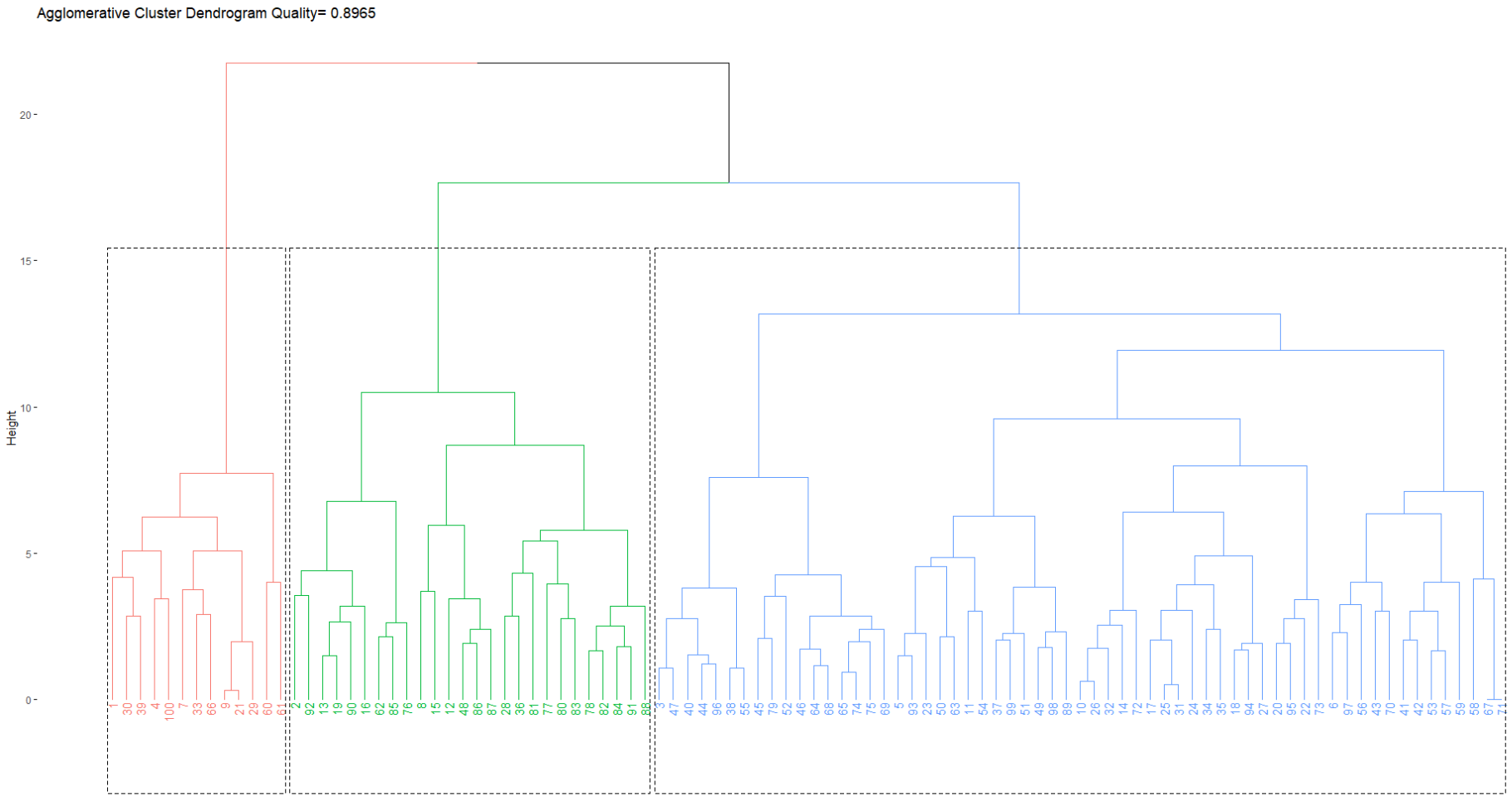


Figure 13 : CAH, Dendrogramme

3.3.3. Détermination de la meilleure partition

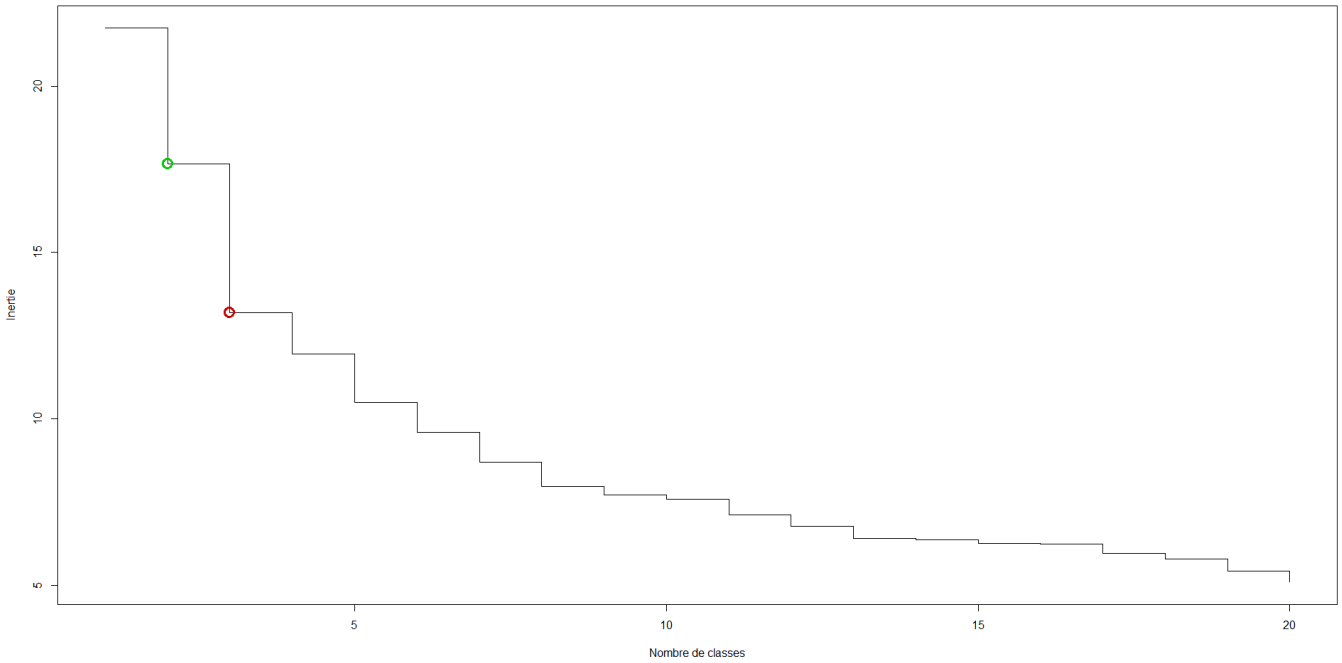


Figure 14: Dissimilarité cumulée en fonction du nombre de classes

Pour déterminer la meilleure partition, on utilise à nouveau la méthode du coude mais en utilisant, cette fois-ci, la dissimilarité cumulée en fonction du nombre de classes. Pour savoir quel est le nombre de classes optimal, on identifie les plus grands sauts de dissimilarité. Dans notre cas, c'est à 2 et 3 classes que l'on obtient les plus grandes pertes de dissimilarité. Nous avons donc considéré ces 2 partitions. Après avoir testé les 2 , nous sommes arrivés à la conclusion que 3 classes semblait être le nombre optimal.

3.3.4. Extraction des profils types :

4. Classification non supervisée :

4.1. Arbres de classifications :

4.1.1. Principes des arbres de classification :

L'apprentissage par arbre de décision est une méthode classique en apprentissage automatique . Son but est de créer un modèle qui prédit la valeur d'une variable-cible depuis la valeur de plusieurs variables d'entrée.

Un arbre se construit de la façon suivante : A chaque nœud de l'arbre, on sélectionne une variable qui sera utilisée comme critère pour diviser la population du nœud en 2. Dans notre exemple, on peut ainsi penser à séparer les individus âgés de plus de 25 ans et de moins de 25

ans. Ce processus est répété ,soit, jusqu'à ce que l'on n'ait plus qu'un seul individu par nœud, ou quand les divisions n'améliorent plus le pouvoir prédictif de l'arbre.

4.1.2. Bootstrap :

Cependant, les arbres sont par nature instables. En effet, l'arbre obtenu dépend fortement du jeu de données sur lequel il a été entrainé. Pour remédier à ce problème, on utilise la technique du Bootstrap. Cette méthode se déroule de la façon suivante :

- Séparer le jeu de données de façon aléatoire en entrainement / validation. Dans notre cas, nous avons, en raison du faible nombre de données, opté pour du 60/40
- Entrainer un arbre et le tester sur le jeu de validation. On note l'erreur (ou précision dans notre cas) obtenue
- On recommence
- On sélectionne l'arbre ayant la meilleure précision

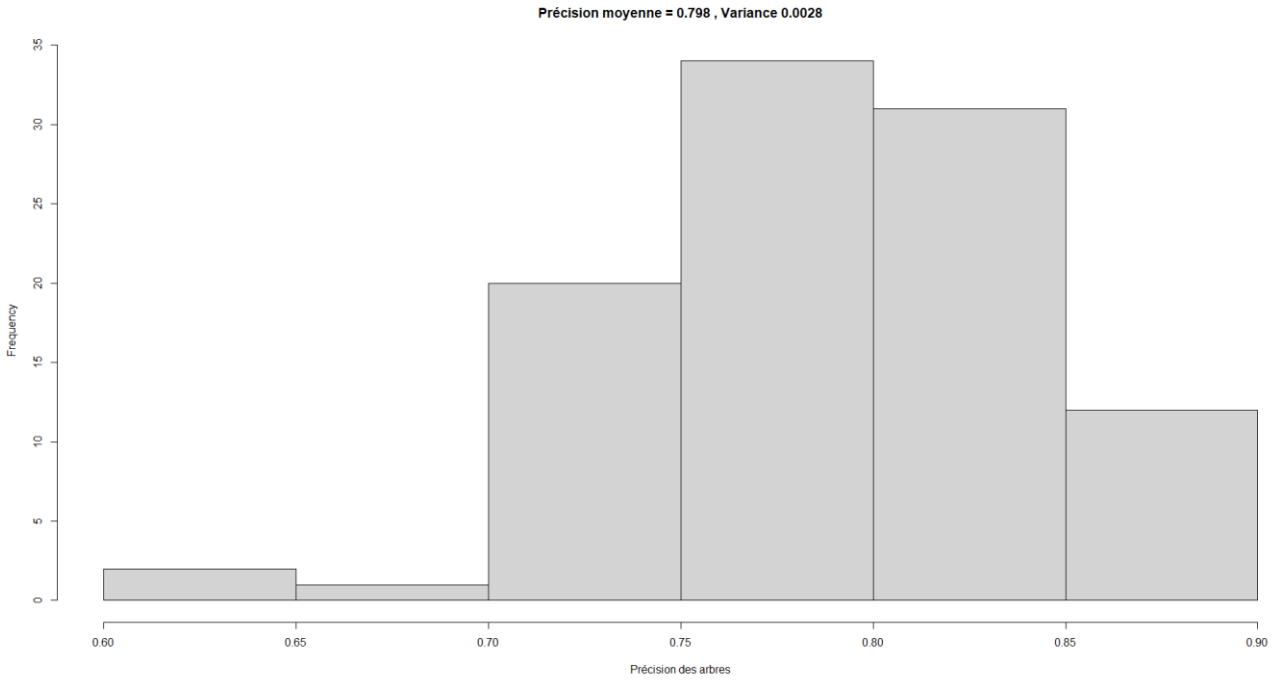


Figure 15: Résultats du Bootstrap

Pour notre jeu de données, nous avons créé 100 arbres et obtenu les résultats ci-dessus. La précision moyenne est de 0.798 et la variance de 0.0028. Il s'agit d'un excellent résultat. En effet, les arbres obtenus ont un bon pouvoir prédictif. De plus, la variance est faible, cela signifie que le modèle est très constant.

Visualisons maintenant l'arbre ayant réalisé la meilleure performance.

4.1.3. Arbre final :

Le meilleur arbre que nous avons obtenu fonctionne de la manière suivante :

- La racine est divisée en appliquant le critère suivant : « Le nombre d'heures passées assis par jour est-il supérieur à 1,52 h (0,095 *16 pour enlever la normalisation). Seul individu ne répond pas à ce critère et son diagnostic est 0
- Les 59 autres individus sont à nouveau divisés selon cette variable mais le seuil est différent : 4,48 h (0.28*16).

On remarque qu'une feuille regroupe 5 individus O (encadrée en rouge). Analysons leur parcours au travers de l'arbre :

- Nombre d'heures passées assis par jour supérieur à 1,52 h
- Nombre d'heures passées assis par jour supérieur à 4,48 h
- Âge supérieur à 23,76 ans (0.66 * 36 ans)
- Ont expérimenté des accidents ou traumatismes
- Test effectué en été ou automne
- Âge inférieur à 28,08 ans (0.78*36)

On observe donc que passer de nombreuses heures assis semble jouer un rôle important dans le diagnostic de fertilité.

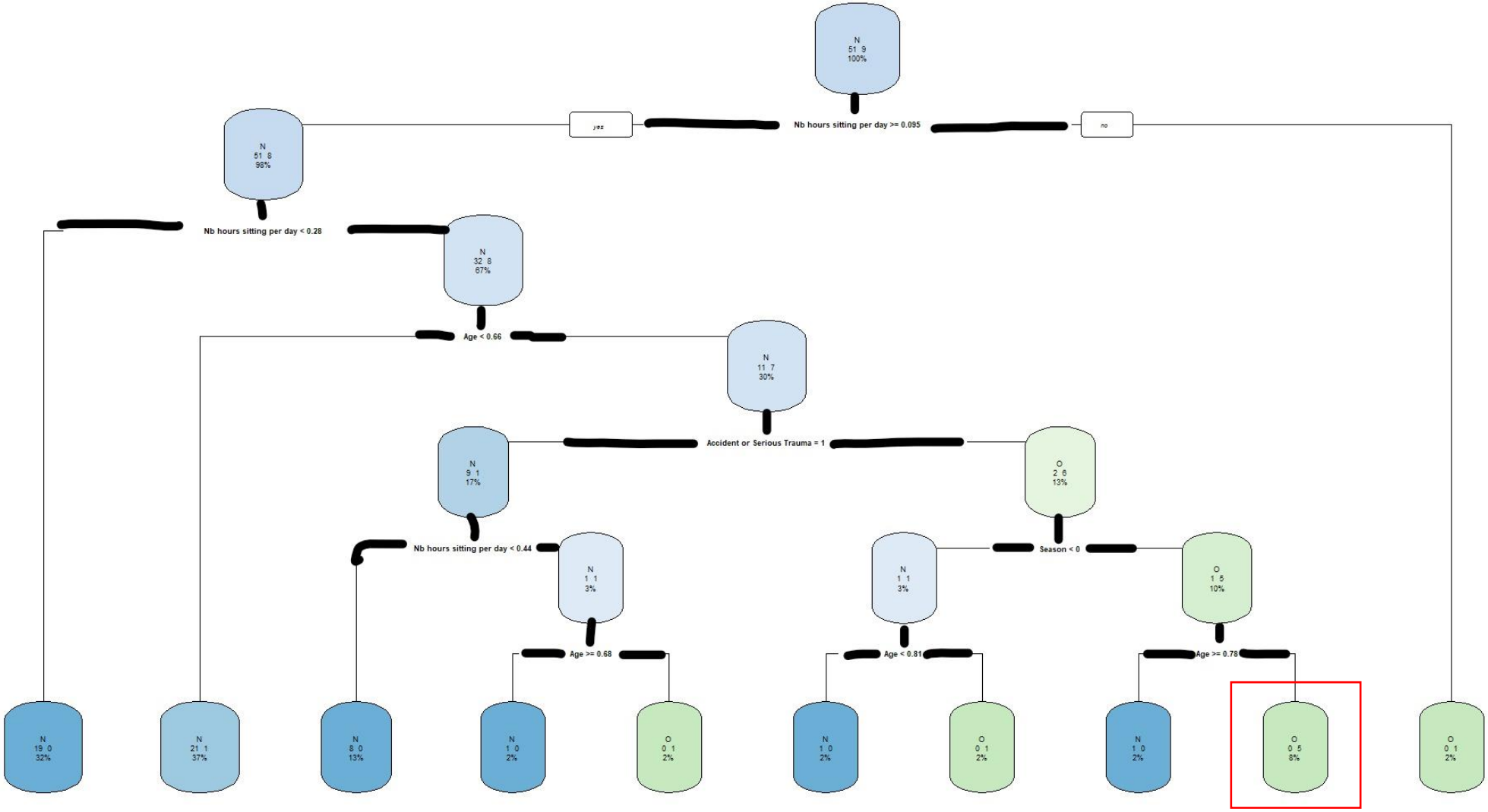


Figure 16 : Arbre de classification final

4.2. Random Forest :

4.2.1. Principe du Random Forest :

Bien que performants, les arbres de classifications présentent quelques inconvénients. En effets, ces deniers peuvent s'avérer instables. De plus, l'arbre final obtenu dépend fortement de l'importance donné à chaque variable. Il est donc plus difficile de généraliser un arbre de classification.

Les forets d'arbres décisionnaires (Random Forest) ont vocation à résoudre ces problèmes. Le principe est simple. Le but est d'entrainer de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. Grâce à cette méthode, toutes les variables sont testées. Par conséquent, il est possible d'obtenir un modèle plus robuste et plus précis.

La contrepartie réside dans le fait qu'un modèle Random Forest n'est pas interprétable de façon globale. En effet, il s'agit désormais d'une multitude d'arbres de classification mélangées. Toutefois, il est possible d'obtenir avec des packages tels que LIME (Local Interpretable Model-Agnostic Explanations)[3]

4.2.2. Modèle :

Pour créer ce modèle Random Forest, nous avons utilisé le package Caret de R et son module « Ranger » en raison de sa compatibilité native avec le package LIME. Nous avons entrainé un modèle avec les paramètres suivants :

- Une cross-validation avec Grid Search répétée 10 fois
- L'impureté de Gini utilisée en tant que critère de qualité des divisions
- 100 arbres

Les paramètres testés dans la Grid Search étaient les suivants :

- Le nombre minimum d'individus dans un nœud, allant de 2 à 10.
- Le nombre de variables choisies à chaque division (mtry), allant de 2 à 9

Une fois le modèle entrainé, nous l'avons testé sur le set d'entrainement puis sur le set de test. Dans les 2 cas le résultat a été satisfaisant. En effet, on obtient des scores de 98,33% et 85% de précision respectivement. Voici les matrices de confusion des tests :

Jeu d' Entrainement	Valeur réelle = N	Valeur réelle = O
Prédiction = N	53	1
Prédiction = O	0	6

Jeu de Test	Valeur réelle = N	Valeur réelle = O
Prédiction = N	33	4
Prédiction = O	2	1

[3]Local Interpretable Model-Agnostic Explanations. https://lime.data-imaginist.com/. Consulté le 23 janvier 2021.

On remarque toutefois que le modèle a du mal à classifier les individus ayant le label O correctement. On peut supposer que c'est dû à la trop faible proportion (12%) d'individus présentant cette caractéristique.

4.2.3. Explication du modèle avec LIME :

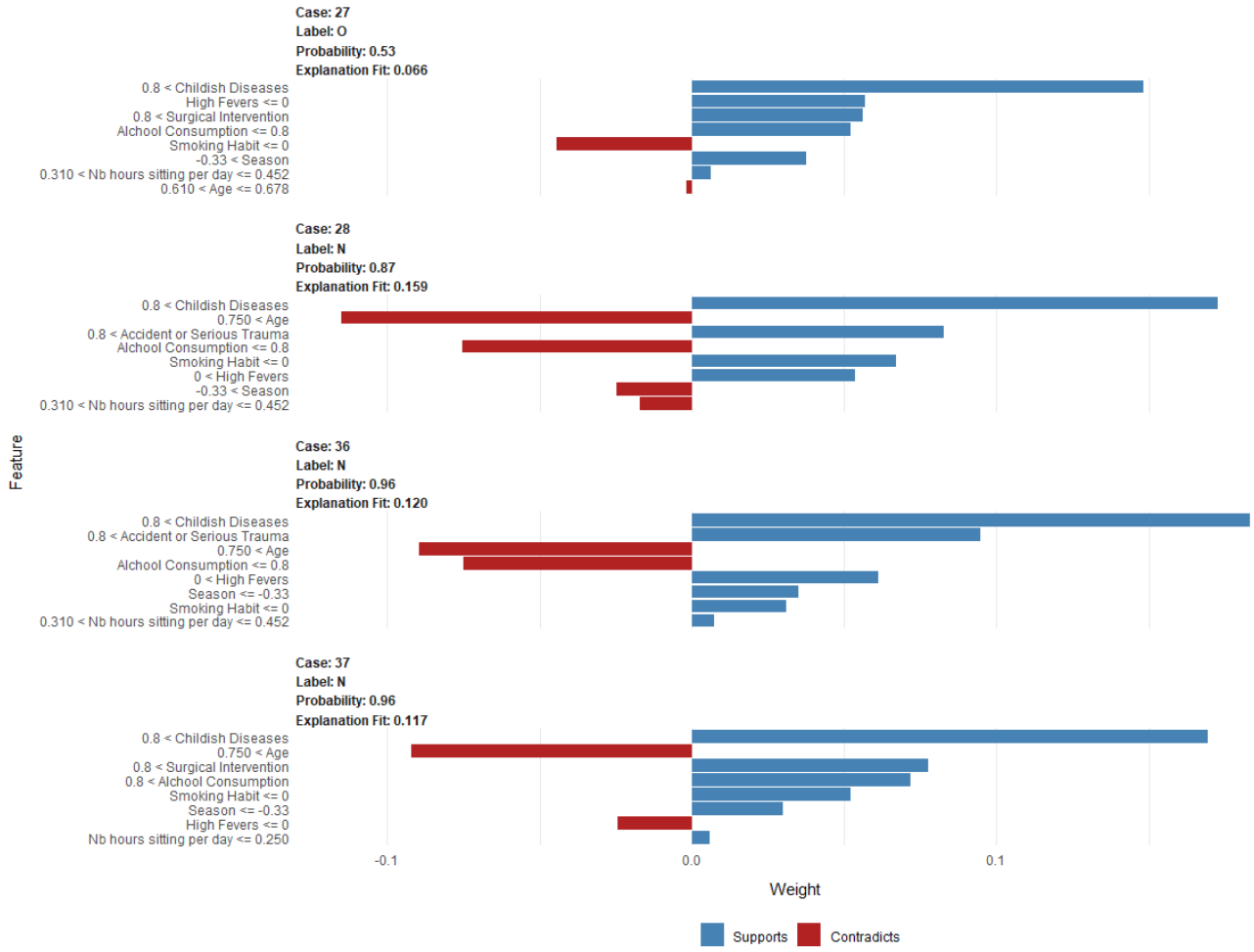


Figure 17: Explication du Random Forest avec LIME

Comme nous l'avons écrit précédemment, il est impossible d'obtenir une interprétation globale d'un modèle Random Forest Cependant, il est possible d'obtenir une interprétation locale avec des packages tels que LIME.

Cela signifie que pour chaque instance, on peut obtenir une description des variables qui ont été sélectionnées par le modèle pour justifier sa prédiction du label. Dans cet exemple, on observe ainsi que pour l'individu n°27, le modèle a prédit à juste titre le label O. Les variables appuyant cette hypothèse sont notamment :

- De fortes fièvres dans les 3 derniers mois
- A eu un accident ou un traumatisme important

A contrario, le fait que l'individu ne fume pas rentre en contradiction avec le label O.

Il est donc possible de d'obtenir pour un individu une interprétation du label qui lui a été prédit.

5. Conclusion :

Les variables épigénétiques étudiées semblent ainsi bel et bien avoir un impact sur la fertilité d'un individu. Cependant, cet impact reste faible au niveau de chaque variable. Par conséquent, il est difficile de prédire avec une grande précision le statut fertile / infertile d'un individu. Les variables semblant jouer le plus dans l'apparition de l'infertilité sont :

- Un âge relativement élevée (+ de 30 ans)
- Une position assise maintenue trop longtemps (+ de 4,5 heures par jour)
- Avoir régulièrement de fortes fièvres
- Avoir eu un accident ou un traumatisme important