Projet de science des données
Fertility data set UCI

Autrices: ASSI Dima & Houda EL AJI
INFO5-Polytech Grenoble

1. Motivation et positionnement du projet

Plusieurs études scientifiques récentes ont noté que la qualité du sperme masculin diminue considérablement en raison du mode de vie et des facteurs environnementaux. Le diagnostic de la qualité du sperme est un aspect important de l'identification du potentiel du sperme pour évaluer la fertilité au sein d'un couple. En raison des progrès des algorithmes d'apprentissage automatique, en particulier de la précision de classification fiable et élevée de la science des données dans différents problèmes de la vie quotidienne, il devient possible de prédire cette qualité à partir des données sur le mode de vie. À cet égard, nous avons procédé par une analyse descriptive et globale, une classification non supervisée et une classification supervisée pour un jeu de données.
Les données utilisées dans cette étude sont collectées à partir du référentiel de données UCI: Fertility Data Set.
Il s'agit de cent (100) volontaires fournissent un échantillon de sperme analysé selon les critères OMS 2010. De ce fait, la base de donnée contient 100 individus (lignes) et 10 attributs qui se présente comme le suivant :

1. Saison au cours de laquelle l'analyse a été effectuée.
    - 1) hiver : -1
    - 2) printemps : -0.33
    - 3) été : 0.33
    - 4) automne : -1
2. Âge au moment de l'analyse : 18-36 (0, 1)
3. Maladies infantiles (c.-à-d. Varicelle, rougeole, oreillons, polio)
    - 1) oui 0
    - 2) non 1
4. Accident ou traumatisme grave
    - 1) oui : 0
    - 2) non :1
5. Intervention chirurgicale
    - 1) oui : 0
    - 2) non : 1

6. Fièvres élevées l'année dernière :
    - 1) il y a moins de trois mois : -1
    - 2) il y a plus de trois mois : 0
    - 3) non : 1
7. Fréquence de la consommation d'alcool : varie en (0, 1)
    - 1) plusieurs fois par jour
    - 2) tous les jours
    - 3) plusieurs fois par semaine
    - 4) une fois par semaine
    - 5) presque jamais ou jamais
8. Fumer :
    - 1) jamais : -1
    - 2) occasionnellement : 0
    - 3) tous les jours. 1
9. Nombre d'heures passées assis par jour: varie entre 1-16 (0, 1)
10. Sortie: Diagnostic
    - normal (N)
    - modifié (O)

2. Analyse descriptive

2.1 Exploration des données

Tout d'abord, nous commençons par faire une exploration des données obtenue par cette étude pour avoir une vue générale de nos individus et leurs habitudes.

En regardant les histogrammes ci-dessous on remarque que la grande majorité des individus ne consomment pas l'alcool fréquemment (cf. Figure 1).

En plus, une grande majorité des individus ~80 passe entre 0 et 8 heure assis durant leur journée(cf. Figure 2).

Et dans le 3ème histogramme, l'âge des individus est entre 0.5 et 1 c'est à dire varie entre 18 et 36 avec la majorité des individus sont âgés entre 18 et 27 ans (cf. Figure 3).

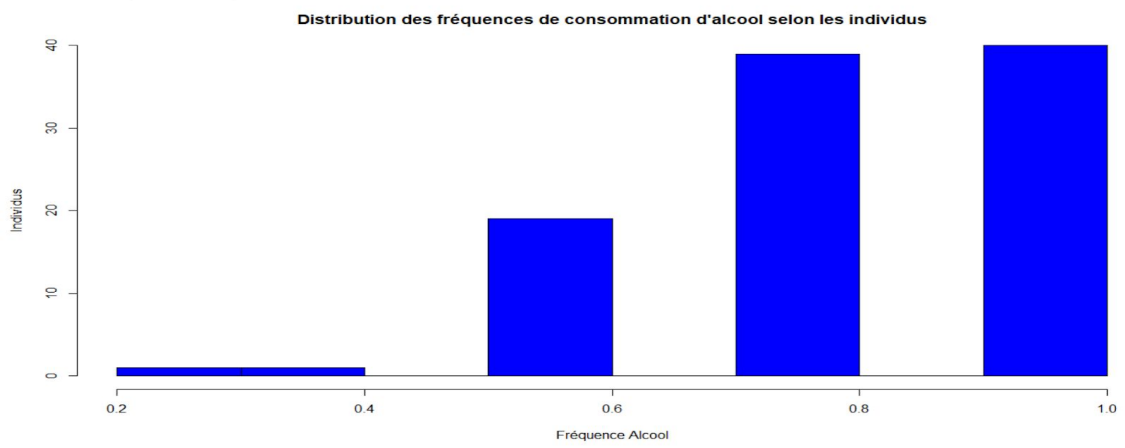


Figure 1. Nombre des individus en fonction de la fréquence de consommation d'alcool

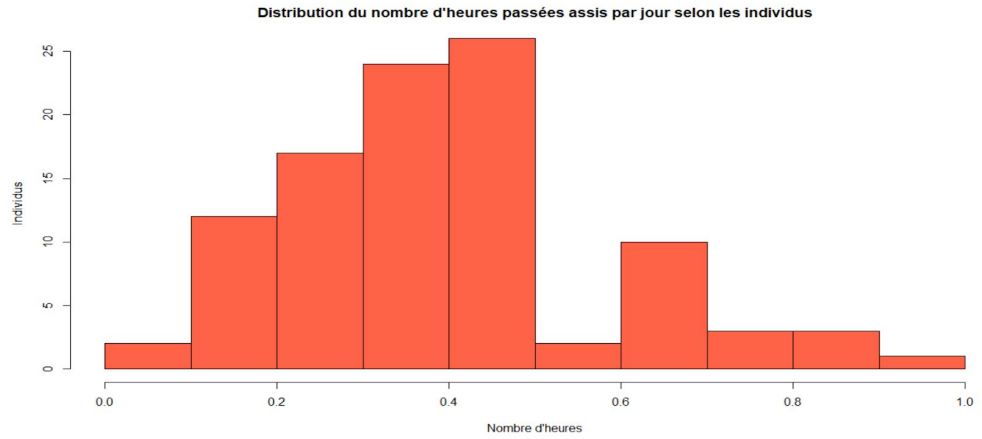


Figure 2. Nombre d'individus en fonction de nombre d'heures passée assis pas jour

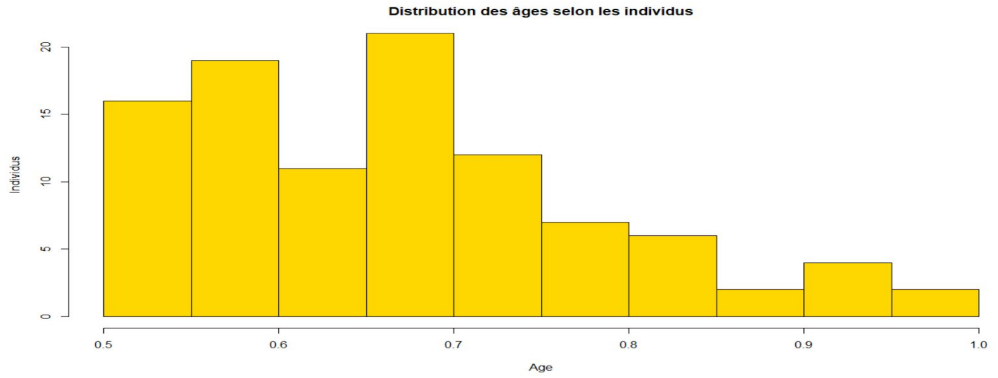


Figure 3. Nombre d'individus en fonction de leurs âges

```{r}
fertilityData[1:10]
summary(fertilityData)
```

data.frame
100 x 10

R Console

Season <dbl>	Age <dbl>	ChildDisease <int>	Trauma <int>	SurgicalInterven <int>	HighFeverLastYear <int>	FrequencyAlcohol <dbl>
-0.33	0.69	0	1	1	0	0.8
-0.33	0.94	1	0	1	0	0.8
-0.33	0.50	1	0	0	0	1.0
-0.33	0.75	0	1	1	0	1.0
-0.33	0.67	1	1	0	0	0.8
-0.33	0.67	1	0	1	0	0.8
-0.33	0.67	0	0	0	-1	0.8
-0.33	1.00	1	1	1	0	0.6
1.00	0.64	0	0	1	0	0.8
1.00	0.61	1	0	0	0	1.0

```
     Season             Age          ChildDisease      Trauma     SurgicalInterven HighFeverLastYear
FrequencyAlcohol
 Min.   :-1.0000   Min.   :0.500   Min.   :0.00   Min.   :0.00   Min.   :0.00     Min.   :-1.00     Min.
:0.200
 1st Qu.:-1.0000   1st Qu.:0.560   1st Qu.:1.00   1st Qu.:0.00   1st Qu.:0.00     1st Qu.: 0.00     1st
Qu.:0.800
 Median :-0.3300   Median :0.670   Median :1.00   Median :0.00   Median :1.00     Median : 0.00
Median :0.800
 Mean   :-0.0789   Mean   :0.669   Mean   :0.87   Mean   :0.44   Mean   :0.51     Mean   : 0.19     Mean
:0.832
 3rd Qu.: 1.0000   3rd Qu.:0.750   3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.00     3rd Qu.: 1.00     3rd
Qu.:1.000
 Max.   : 1.0000   Max.   :1.000   Max.   :1.00   Max.   :1.00   Max.   :1.00     Max.   : 1.00     Max.
:1.000
    Smoking       SittingHour     Diagnosis
 Min.   :-1.00   Min.   :0.0600   N:88
 1st Qu.:-1.00   1st Qu.:0.2500   O:12
 Median :-1.00   Median :0.3800
 Mean   :-0.35   Mean   :0.4068
 3rd Qu.: 0.00   3rd Qu.:0.5000
 Max.   : 1.00   Max.   :1.0000
```

Nos descripteurs comme données et comme vous pouvez les voir dans les résultats ci-dessus sont:

- ***Season***, ***HighFeverLastYear***, ***Smoking***: Variables qualitatives ordinales
- ***ChildDisease, Trauma, SurgicalIntervention, Diagnosis***: Variables binaires
- ***Age, FrequencyAlcohol, SittingHour:*** Variables numériques

Les résultats obtenus pour faire l'analyse descriptive montrent que parmi les 100 individus, 88 ont des résultats normaux de concentration de sperme tandis que le reste(12 individus) présentent des résultats altérés.

L'âge moyen des individus est 0.669 qui est équivalent à ~24 ans.

2.2 Analyse des relation entre les attributs

2.2.1 Base de projection

Sur la base de projection ci-dessous nous pouvons mettre plus en valeur les relations entre les instances il ya un grand nombre qui n'ont pas de relations claires ou d'effets l'un sur l'autre comme Age-ChildDisease Age-Trauma ChildDisease-Trauma ce qui est normal tandis qu'on voit bien que Age-Sitting hour ont une relation triangulaire.
En plus on voit pas d'effet directe claire de Smoking, HighFeverLastYear, ChildDisease, Season,Trauma, ChirugicalInteventionsur Diagnosis

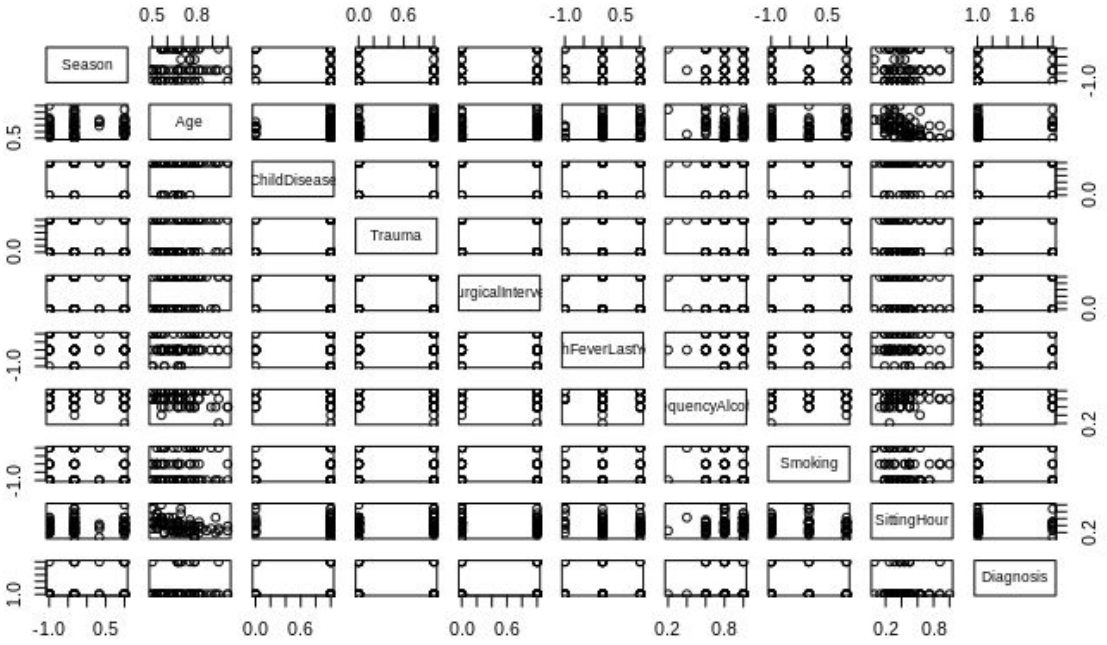
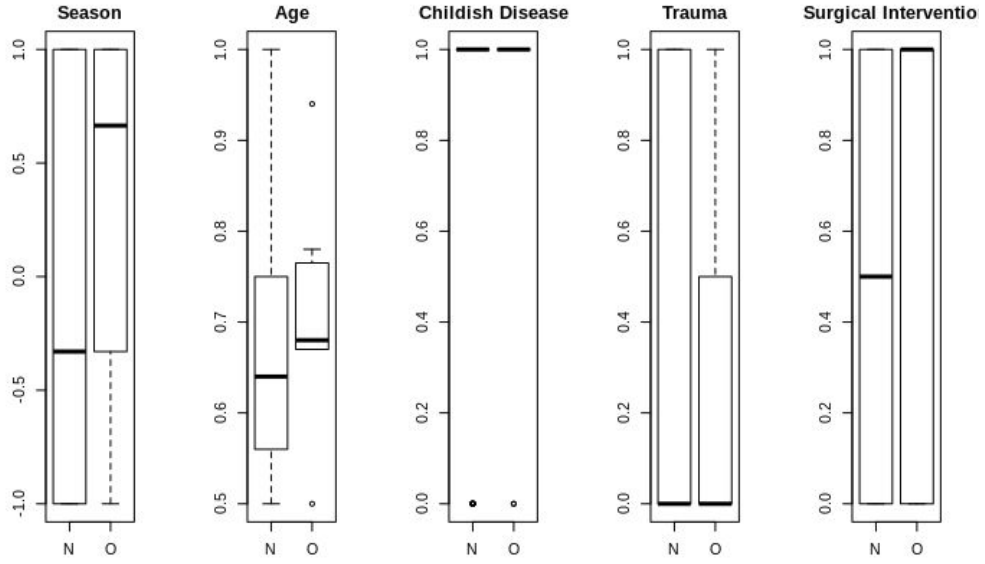


Figure 4. Base de projection

2.2.2 Analyse des boxplots

Voici les boxplots de notre étude, pour chaque instance nous avons créé deux boxplots, une pour les individus ayant des diagnostics normaux et une pour les individus ayant des diagnostics altérés.



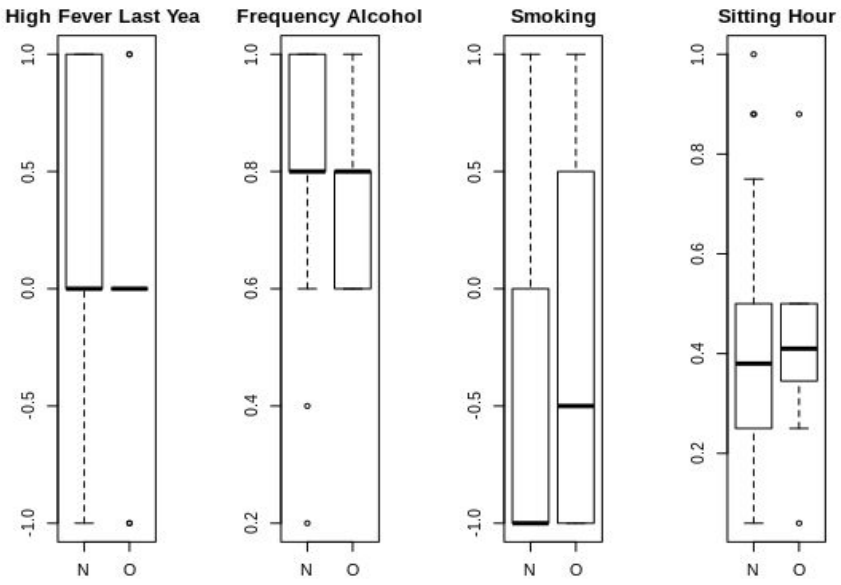


Figure 5.Boxplots

Age: L'âge maximal des individus avec des résultats altérés est de ~0.79 ce qui est équivalent à ~29 avec une moyenne de 0.69 qui est équivalent à ~25.
Le nombre d'heure passé assis pour les individus normales est 0 minimum tandis que les individus avec des résultats altérés
Pareil, vous pouvez voir tous les résultats des autres instances dans la figure 4.

2.2.3 Analyse de relations entre l'âge et le résultat de diagnostics

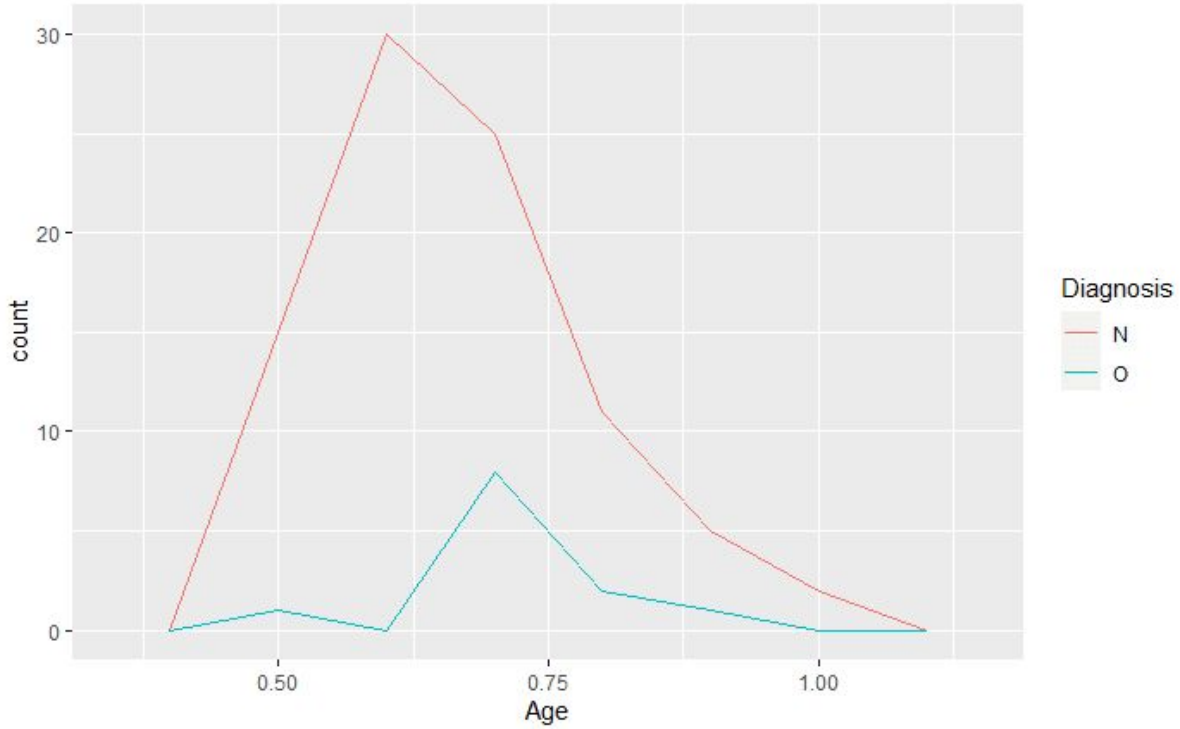


Figure 6. Résultat de diagnostics en fonction de l'âge

Dans la courbe ci-dessus on examine l'influence de l'âge sur les diagnostics, on remarque que dans l'intervalle d'âge donné dans l'étude le nombre des individus avec des résultats normaux est beaucoup plus important que ceux qui ont des résultats altérés et c'est normalement dû au fait que les individus dans cet intervalle entre 18 et 36 ans sont dans une période de leur vie où leur fertilité doit être la meilleur.

2.2.4 Analyse de relation entre la fréquence de consommation d'alcool et les saisons

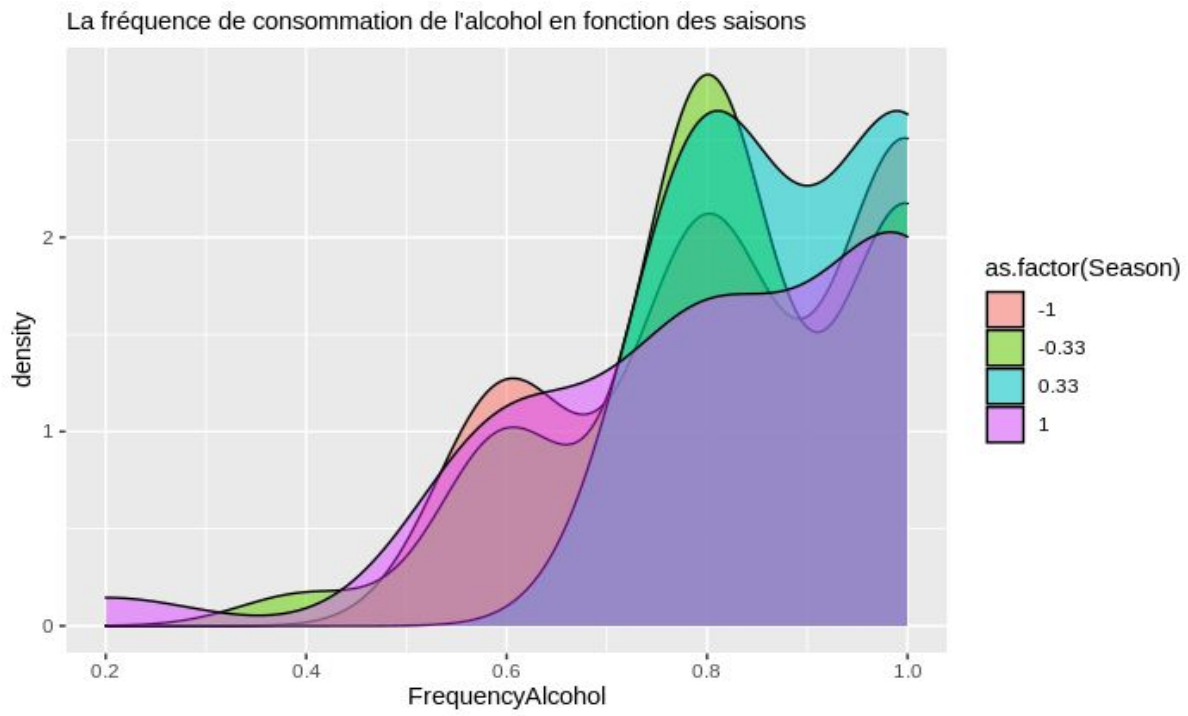


Figure 7.Fréquence de consommation d'alcool en fonction des saisons

Les courbes ci-dessus présentent la fréquence de consommation d'alcool en fonction des saisons, on remarque que durant l'été est la période la plus importante au niveau de consommation d'alcool, suivi du printemps puis l'hiver et finalement l'automne ce qui est normal concernant les individus qu'on a dans cette étude.

2.3 Conclusion

Jusqu'à là et après avoir une vue globale des relations entre les attributs on ne peut pas avoir des analyses claires et définies de ce qui affectent le plus les résultats des diagnostics mais d'autre part on peut voir que certains attributs ont des effets directs sur les résultats. En plus certains diagrammes qu'on a mis dans nos analyses montre clairement des relations directe entre les attributs comme ***season*** et ***AlcoholFrequency...***

3. Classification non supervisée

3.1. Matrice de dissimilarités :

Ayant des attributs de type mixte (c-à-d : attributs binaire, ordinale , numérique), nous avons opté pour l'utilisation de la méthode Gower. Le concept de distance de Gower est en fait assez simple. Pour chaque type de variable, une mesure de distance particulière qui

fonctionne bien pour ce type est utilisée et mise à l'échelle pour tomber entre 0 et 1. Ensuite, une combinaison linéaire utilisant des poids spécifiés par l'utilisateur (plus simplement une moyenne) est calculée pour créer la matrice de distance finale . Les métriques utilisées pour chaque type de données sont décrites ci-dessous:

1. quantitatif (intervalle): distance de Manhattan normalisée par distance
2. ordinal: la variable est classée en premier, puis la distance de Manhattan est utilisée avec un ajustement spécial pour les liens
3. nominal: les variables de k catégories sont d'abord converties en k colonnes binaires puis le coefficient de dés est utilisé.

3.2 Classification ascendante hiérarchique

3.2.1 Arbre de CAH

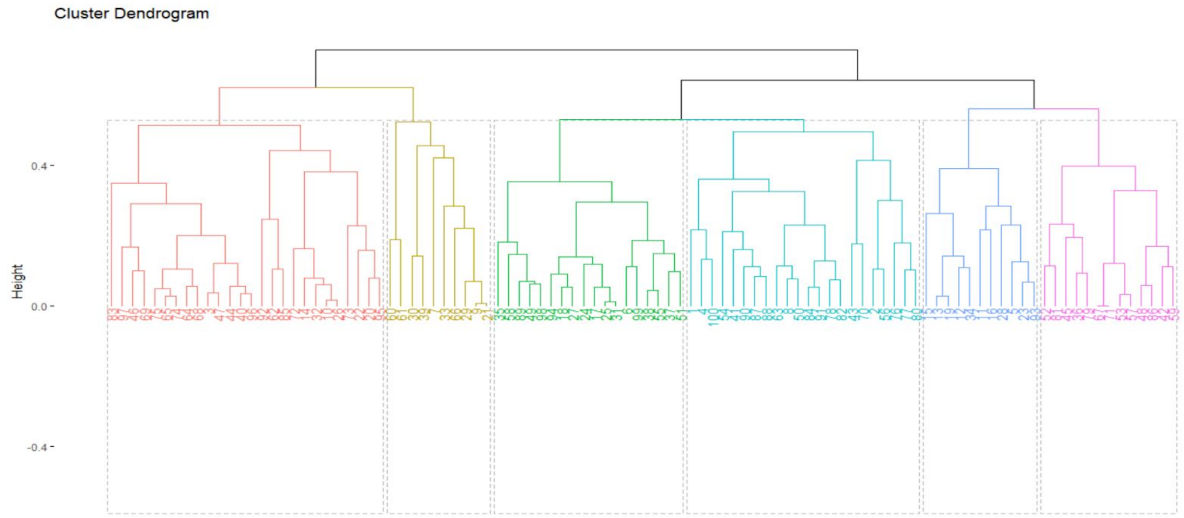


Figure 8. Arbre CAH

La dessous on observe un dendrogramme où l'on peut choisir de retenir 6 classes avec une moyenne de similarité d'environ 0.6 qui est une valeur qu'on trouve intéressante et qu'on a trouvé la meilleure parmi tous les découpages de classes qu'on avait testé.

3.3 Partitionnement autour des medoids (PAM)

3.3.1 Choix de nombre de cluster:

Dans notre étude nous avons choisi d'avoir 6 clusters pour vérifier ce choix au lieu d'un nombre plus petit ou plus grand nous avons ajouté la courbe ci-dessous qui montre que 6 est le meilleur choix pour avoir la meilleur largeur de silhouette

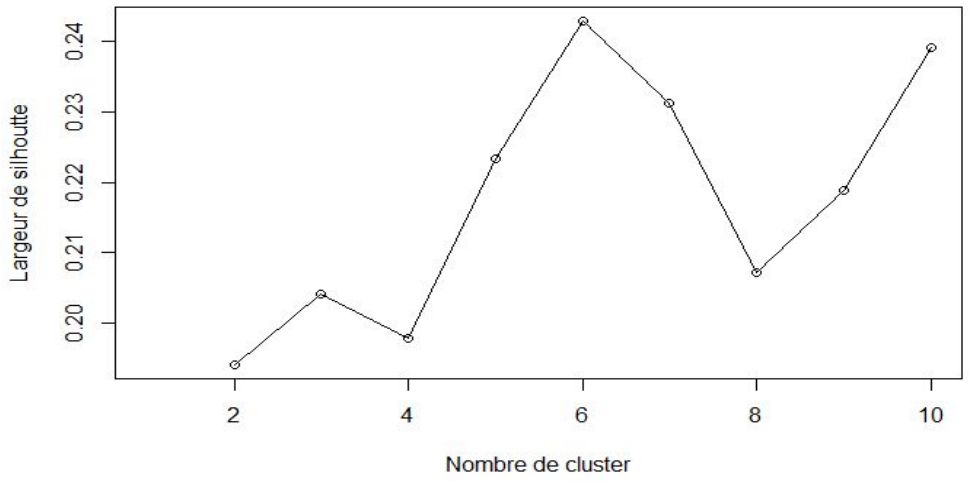


Figure 9. Choix de nombre de cluster

3.3.2 Analyse silhouette

On observe dans la figure ci-dessous 6 classes différentes, on voit clairement que surtout dans la première plein d'individus sont mal classés et elle a une moyenne Si très basse.
La deuxième a des individus classés de la meilleure façon, on voit qu'un seul est mal classé et la valeur moyenne Si dans cette classe est la plus proche de 1, elle est la plus stable.
On peut voir clairement que dans chaque bloc d'individus il y a 1 ou plusieurs individus mal classés et la meilleure moyenne de valeur Si est de 0.41.

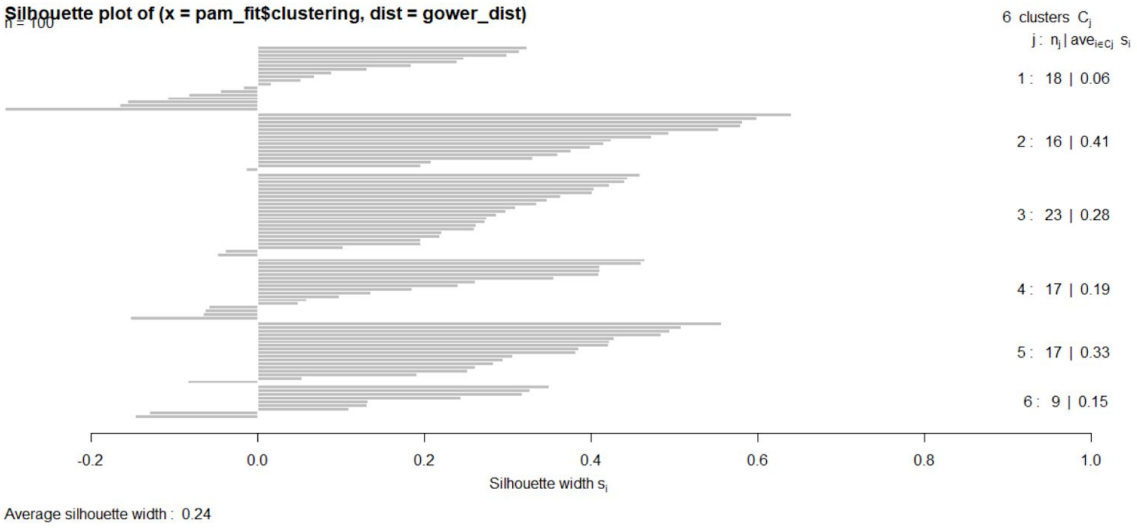


Figure 10.Silhouette

3.3.2 Analyse des clusters obtenus et profils types :

Visualisation

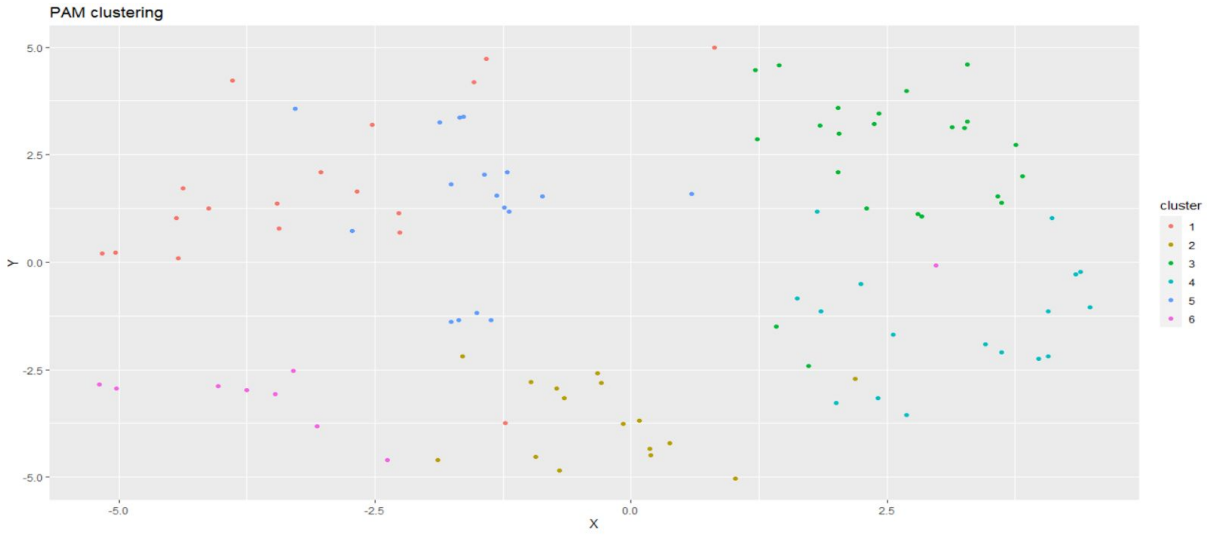


Figure 11.PAM

Après avoir exécuté l'algorithme PAM et sélectionné six clusters, nous pouvons interpréter les clusters en exécutant un résumé sur chaque cluster (cf Figure ci-dessous). Sur la base de ces résultats, il semble que

- Le cluster 1 : Les individus dans ce groupe ont généralement passé le test entre hiver et printemps, et ont une moyenne d'âge de 0.675, 75% d'eux n'avaient pas des Child Disease tandis qu'une petite partie d'eux a eu des traumas, et peuvent être considérés comme alcoolique.
- Le cluster 2 :Dans ce groupe l'âge moyen est 0.6215 et aucun des individus n'a eu des Child disease (min = 1) et ils ont tous eu des intervention chirurgical(max = 0)
- Le cluster 3 :Ce groupe a un âge moyen de 0.7609, aucun individu n'a eu des Traumas(min = 1) et les individus de ce groupe ont relativement alcoolique.
- Le cluster 4 :Ce groupe a 0.67 comme âge moyen, et aucun des individus n'a eu des Child disease ni des Traumas et la grande majorité n'a pas eu d' interventions chirurgicales.
- Le cluster 5 :La grande majorité de ce groupe a eu des traumas, les individus de ce groupe sont généralement des consommateurs d'alcool, mais ne sont pas fumeurs(max = -1)
- Le cluster 6 :La majorité de ce groupe ont eu des Child diseases, la majorité ont eu des traumatismes, et tous les individus de ce groupe ont eu des interventions chirurgicales.

```
[[1]]
    Season            Age          ChildDisease        Trauma        SurgicalInterven HighFeverLastYear FrequencyAlcohol
 Min.   :-1.0000   Min.   :0.5000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :-1        Min.   :0.6000
 1st Qu.:-1.0000   1st Qu.:0.5875   1st Qu.:0.2500   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.: 0        1st Qu.:0.8000
 Median :-0.3300   Median :0.6700   Median :1.0000   Median :0.00000   Median :1.0000   Median : 0        Median :0.8000
 Mean   :-0.5539   Mean   :0.6750   Mean   :0.7222   Mean   :0.05556   Mean   :0.8889   Mean   : 0        Mean   :0.8111
 3rd Qu.:-0.3300   3rd Qu.:0.6850   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.: 0        3rd Qu.:0.9500
 Max.   : 1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   : 1        Max.   :1.0000
    Smoking        SittingHour       cluster
 Min.   :-1.0000   Min.   :0.1900   Min.   :1
 1st Qu.:-1.0000   1st Qu.:0.2650   1st Qu.:1
 Median : 0.0000   Median :0.5000   Median :1
 Mean   :-0.1111   Mean   :0.4489   Mean   :1
 3rd Qu.: 0.0000   3rd Qu.:0.5450   3rd Qu.:1
 Max.   : 1.0000   Max.   :0.8800   Max.   :1

[[2]]
    Season            Age          ChildDisease      Trauma       SurgicalInterven HighFeverLastYear FrequencyAlcohol    Smoking
 Min.   :-1.0000   Min.   :0.5000   Min.   :1     Min.   :0.0000   Min.   :0        Min.   :0.0000    Min.   :0.600    Min.   :-1.00
 1st Qu.:-1.0000   1st Qu.:0.5225   1st Qu.:1     1st Qu.:0.0000   1st Qu.:0        1st Qu.:0.0000    1st Qu.:0.800    1st Qu.:-1.00
 Median :-1.0000   Median :0.5600   Median :1     Median :0.0000   Median :0        Median :1.0000    Median :1.000    Median :-1.00
 Mean   :-0.6656   Mean   :0.6125   Mean   :1     Mean   :0.0625   Mean   :0        Mean   :0.6875    Mean   :0.925    Mean   :-0.75
 3rd Qu.:-0.3300   3rd Qu.:0.6750   3rd Qu.:1     3rd Qu.:0.0000   3rd Qu.:0        3rd Qu.:1.0000    3rd Qu.:1.000    3rd Qu.:-1.00
 Max.   : 0.3300   Max.   :0.9200   Max.   :1     Max.   :1.0000   Max.   :0        Max.   :1.0000    Max.   :1.000    Max.   : 1.00
  SittingHour       cluster
 Min.   :0.1900   Min.   :2
 1st Qu.:0.3100   1st Qu.:2
 Median :0.4400   Median :2
 Mean   :0.4288   Mean   :2
 3rd Qu.:0.5000   3rd Qu.:2
 Max.   :0.6300   Max.   :2

[[3]]
    Season            Age          ChildDisease       Trauma  SurgicalInterven HighFeverLastYear FrequencyAlcohol    Smoking
 Min.   :-1.0000   Min.   :0.5800   Min.   :0.000   Min.   :1   Min.   :0.000    Min.   :-1.0000   Min.   :0.2000   Min.   :-1.0000
 1st Qu.:-0.6650   1st Qu.:0.7050   1st Qu.:1.000   1st Qu.:1   1st Qu.:1.000    1st Qu.: 0.0000   1st Qu.:0.6000   1st Qu.:-1.0000
 Median :-0.3300   Median :0.7500   Median :1.000   Median :1   Median :1.000    Median : 0.0000   Median :0.8000   Median :-1.0000
 Mean   :-0.2157   Mean   :0.7609   Mean   :0.913   Mean   :1   Mean   :0.913    Mean   : 0.2174   Mean   :0.7652   Mean   :-0.4348
 3rd Qu.:-0.3300   3rd Qu.:0.8100   3rd Qu.:1.000   3rd Qu.:1   3rd Qu.:1.000    3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.: 0.0000
 Max.   : 1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1   Max.   :1.000    Max.   : 1.0000   Max.   :1.0000   Max.   : 1.0000
  SittingHour       cluster
 Min.   :0.1300   Min.   :3
 1st Qu.:0.2200   1st Qu.:3
 Median :0.2500   Median :3
 Mean   :0.3009   Mean   :3
 3rd Qu.:0.3800   3rd Qu.:3
 Max.   :0.7500   Max.   :3

[[4]]
    Season            Age          ChildDisease      Trauma  SurgicalInterven HighFeverLastYear FrequencyAlcohol    Smoking
 Min.   :-1.0000   Min.   :0.5000   Min.   :1     Min.   :1   Min.   :0.0000   Min.   :-1        Min.   :0.4000   Min.   :-1
 1st Qu.:-0.3300   1st Qu.:0.5300   1st Qu.:1     1st Qu.:1   1st Qu.:0.0000   1st Qu.: 0        1st Qu.:0.8000   1st Qu.:-1
 Median :-0.3300   Median :0.5600   Median :1     Median :1   Median :0.0000   Median : 0        Median :0.8000   Median : 0
 Mean   :-0.2141   Mean   :0.6029   Mean   :1     Mean   :1   Mean   :0.1176   Mean   : 0        Mean   :0.7765   Mean   : 0
 3rd Qu.:-0.3300   3rd Qu.:0.6700   3rd Qu.:1     3rd Qu.:1   3rd Qu.:0.0000   3rd Qu.: 0        3rd Qu.:0.8000   3rd Qu.: 1
 Max.   : 1.0000   Max.   :0.8900   Max.   :1     Max.   :1   Max.   :1.0000   Max.   : 1        Max.   :1.0000   Max.   : 1
  SittingHour       cluster
 Min.   :0.1900   Min.   :4
 1st Qu.:0.3100   1st Qu.:4
 Median :0.5000   Median :4
 Mean   :0.4976   Mean   :4
 3rd Qu.:0.6300   3rd Qu.:4
 Max.   :0.8800   Max.   :4

[[5]]
    Season            Age          ChildDisease        Trauma        SurgicalInterven HighFeverLastYear  FrequencyAlcohol    Smoking
 Min.   :-1.0000   Min.   :0.5600   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :-1.00000   Min.   :0.6000   Min.   :-1
 1st Qu.: 1.0000   1st Qu.:0.6100   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:0.8000   1st Qu.:-1
 Median : 1.0000   Median :0.6400   Median :1.0000   Median :0.00000   Median :1.0000   Median : 0.00000   Median :1.0000   Median :-1
 Mean   : 0.8429   Mean   :0.6547   Mean   :0.7647   Mean   :0.05882   Mean   :0.7059   Mean   :-0.05882   Mean   :0.9294   Mean   :-1
 3rd Qu.: 1.0000   3rd Qu.:0.6900   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.: 0.00000   3rd Qu.:1.0000   3rd Qu.:-1
 Max.   : 1.0000   Max.   :0.8100   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   : 0.00000   Max.   :1.0000   Max.   :-1
  SittingHour       cluster
 Min.   :0.1900   Min.   :5
 1st Qu.:0.2500   1st Qu.:5
 Median :0.3800   Median :5
 Mean   :0.3924   Mean   :5
 3rd Qu.:0.4400   3rd Qu.:5
 Max.   :0.6300   Max.   :5

[[6]]
    Season            Age          ChildDisease        Trauma       SurgicalInterven HighFeverLastYear FrequencyAlcohol    Smoking
 Min.   :-0.3300   Min.   :0.5600   Min.   :0.0000   Min.   :0.0000   Min.   :0        Min.   :0.0000    Min.   :0.6      Min.   :0.0000
 1st Qu.: 1.0000   1st Qu.:0.5800   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0        1st Qu.:0.0000    1st Qu.:0.6      1st Qu.:0.0000
 Median : 1.0000   Median :0.6700   Median :1.0000   Median :0.0000   Median :0        Median :0.0000    Median :0.8      Median :1.0000
 Mean   : 0.7778   Mean   :0.6744   Mean   :0.7778   Mean   :0.1111   Mean   :0        Mean   :0.4444    Mean   :0.8      Mean   :0.6667
 3rd Qu.: 1.0000   3rd Qu.:0.7800   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0        3rd Qu.:1.0000    3rd Qu.:1.0      3rd Qu.:1.0000
 Max.   : 1.0000   Max.   :0.8100   Max.   :1.0000   Max.   :1.0000   Max.   :0        Max.   :1.0000    Max.   :1.0      Max.   :1.0000
  SittingHour      cluster
 Min.   :0.06   Min.   :6
 1st Qu.:0.25   1st Qu.:6
 Median :0.44   Median :6
 Mean   :0.41   Mean   :6
 3rd Qu.:0.50   3rd Qu.:6
 Max.   :1.00   Max.   :6
```

3.3.3 Kmeans

Pour l'analyse des données qu'on a nous n'avons pas pu faire du kmeans car la grande majorité des variables qu'on a ne sont pas des valeurs numériques.

3.4 Conclusion

A la fin de cette partie nous avons nos individus groupés dans 6 groupes, chaque groupe a un cluster.
Pas tous les individus sont bien classés avec le cluster le plus proche mais on voit des ressemblances claires entre les individus d'un même groupe (nous pensons que cela peut être dû aussi à la nature des attributs de notre dataset, étant dans la majorité qualitative).
Nous avons aussi un choix de nombres de clusters pour avoir les individus classés de la meilleure façon.

4. Classification supervisée

4.1 Arbres de décision

La dessous nous avons mis 2 arbres de décision, une globale de Diagnosis en fonction de toutes les variables(la première) et une présentant ***Diagnosis*** en fonction de ***Trauma***, ***SurgicalIntervention*** et ***Age***.
Dans le second arbre on voit clairement qu'un individu âgé entre 0.66 et 0.77 et ayant une intervention chirurgicale et n'ayant pas expérimenté une **trauma** a une grande chance d'avoir des résultats anormaux de ses **diagnostics**.
Dans le premier, on voit clairement qu'un individu âgé de moins de 0.66 ~24 ans, il y a une grande chance que ces résultats soient normaux.

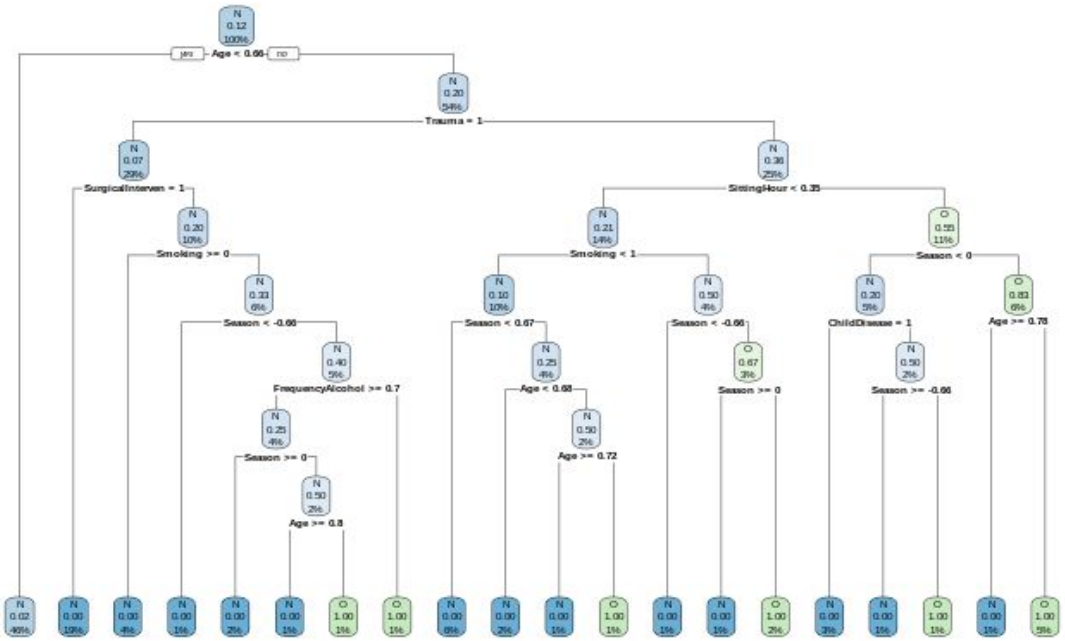


Figure 12.Arbre de décision global

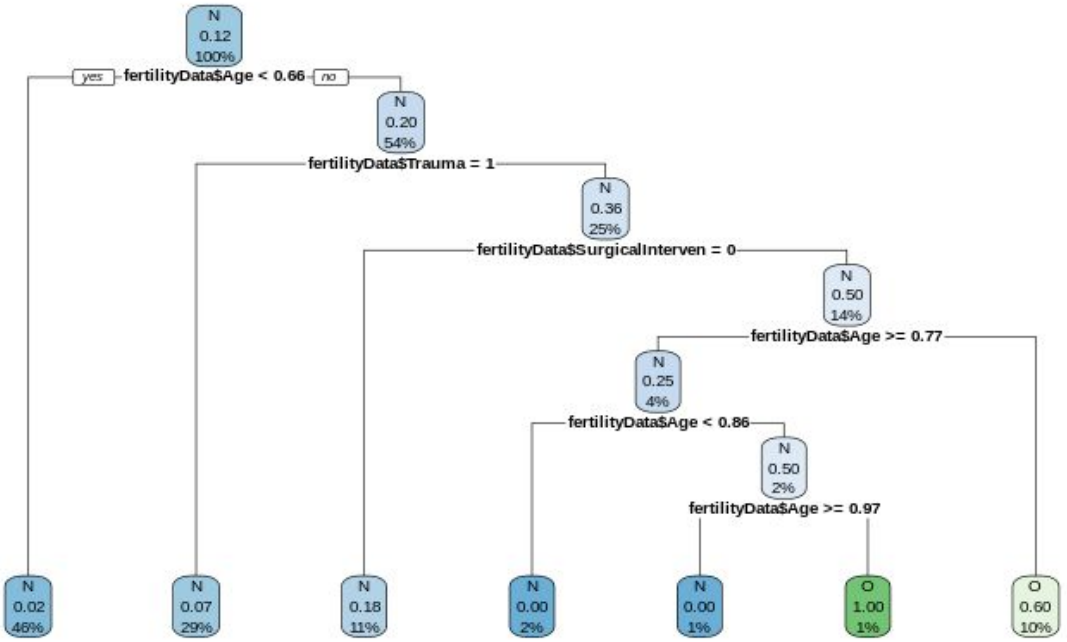


Figure 13.Arbre de décision Age, Trauma, SurgicalIntervention

4.2 Validation croisée

Afin de valider notre arbre de décision et évaluer les erreurs de prédictions nous avons choisi d'utiliser la méthode de validation croisée que nous jugeons plus convenable à notre dataset qui ne contient que 100 individus (taille relativement petite).
On s'est intéressé tout d'abord à calculer l'erreur globale de prédiction des deux arbres obtenus ci-dessus. Pour calculer cette erreur nous avons calculer la **matrice de confusion**

```
#compute global error rate of our first tree and second
pred = predict(arbre1, fertilityData, type = "class")
mc =table(fertilityData$Diagnosis, pred)
print(mc)
t=(mc[1, 2]+mc[2,1])/sum(mc)
print(t)
```

```
   pred
     N  O
  N 84  4
  O  5  7
[1] 0.09
   pred2
     N  O
  N 88  0
  O  1 11
[1] 0.09
```

Ici, pred et pre2 constitue respectivement les matrices de confusion pour arbre global et arbre (Age, Trauma, SurgicalIntervention).
Ainsi, nous remarquons que les deux arbre ont un taux d'erreur semblable et très léger (=0.09)

Afin d'avoir un indicateur d'erreur plus solide, on applique la méthode de validation croisée. Ainsi, nous proposons de sélectionner 70 individus au hasard dans le jeu de données, pour un échantillon d'apprentissage et le reste constitue le jeu de données de test. Puis, on utilise les individus non sélectionnés pour construire le jeu de données d'apprentissage avec lequel on va construire un nouvel arbre de classification (cf. code source), dont on calcule le taux d'erreur apprentissage-classification. L'arbre est sélectionné en minimisant un critère d'erreur (estimé par validation croisée) en fonction d'un paramètre appelé `cp`.

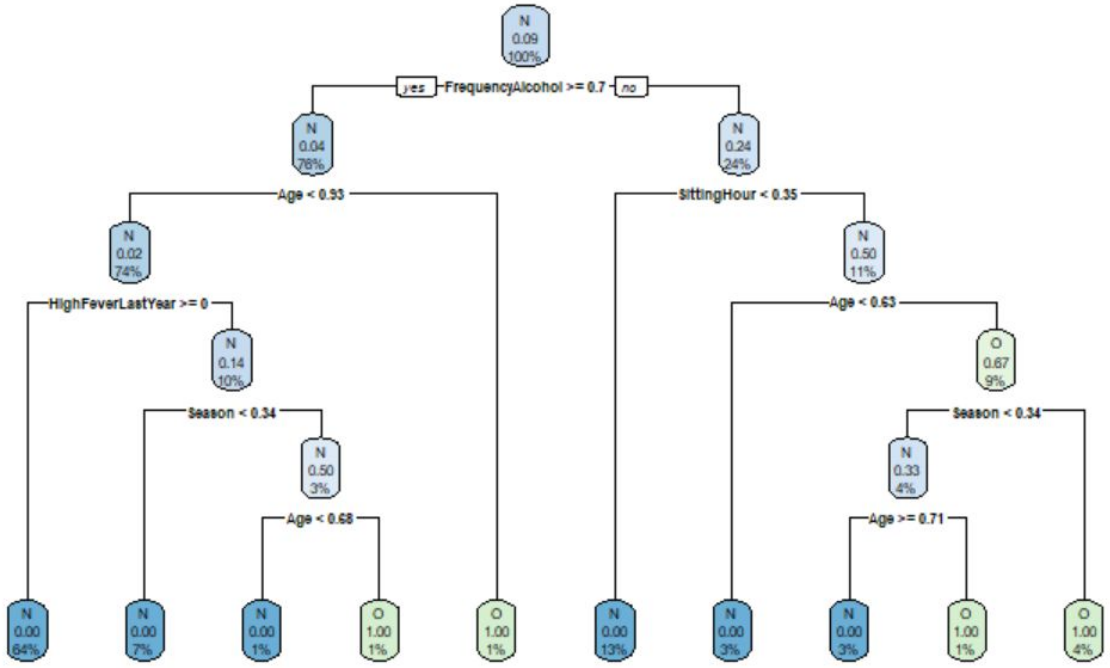


Figure 14. Arbre obtenue avec méthode de validation croisée

L'arbre obtenu est différent des deux premiers, avec une valeur d'erreur égale à 20% environ. Nous constatons clairement que les performances de cet arbre est plus faible que ce que nous avons construit auparavant.Nous voyons par ailleurs que les valeurs de `cp` prises par défaut ne permettent pas de minimiser le critère d'erreur. Il faut augmenter la plage de valeurs de `cp` sur laquelle la minimisation est calculée.

4.3 Conclusion :

D'après cette étude de classification supervisée nous constatons qu'il y a des facteurs qui affectent directement les résultats par exemple toutes les personnes âgées de moins que 33 ans (0.66) ont des résultats normaux.
En plus nous trouvons aussi que plein d'autres facteurs combinés peuvent avoir un effet sur les résultats de diagnostics comme par exemple avoir eu des SurgicalIntervention et des Traumas.
Finalement la validation croisée a validé les résultats de l'arbre de la figure 13.

5. Conclusion globale

Dans ce rapport nous avons fait nos analyses en utilisant plusieurs classifications de données. Nous pouvons conclure que certaines valeurs ont joué un rôle plus important que d'autres en affectant le résultat des données et pour prédire le diagnostic du sperme . En effet, Il y avait beaucoup de valeurs qui avaient peu ou pas d'impact sur le résultat du test de fertilité, comme la saison, le temps passé assis. Pourtant, l'âge à titre d'exemple peut impacter la qualité séminale. Nous pouvons conclure alors que les facteurs liés à la santé des hommes et leur antécédent sont considérés plus importants que leurs habitudes et l'environnement dans lequel ils vivent.

Finalement, nous considérons que ces résultats peuvent toujours être améliorés avec un échantillon plus grand, car plus on a d' individus, plus la véracité de notre analyse et généralisation de synthèse est meilleure.