Rapport de Analyse de Donnée

Cencus Incomes

Etudiant :
Guillaume NGUYEN Main 4
Andrei FLEISER Main 4

Enseignant :
Fanny VILLERS

Année 2018-2019

Table des matières

1 Introduction **2**

2 Le jeu de données **2**

3 Réduction de dimensions **2**

3.1 ACM . . . 2

3.2 FAMD . . . 3

4 Apprentissage non-supervisé **5**

4.1 K-means sur 2 classes . . . 5

4.2 K-means sur plusieurs classes. . . . 6

4.3 CAH . . . 7

4.3.1 Résultat et intérprètation . . . 7

4.3.2 Comparaison avec le K-means . . . 9

5 Apprentissage supervisé **9**

5.1 CART . . . 9

5.1.1 Arbre de classification standard . . . 9

5.1.2 Arbre optimal . . . 11

5.2 RANDOM FOREST . . . 11

5.3 Régression logistique . . . 12

5.3.1 Tests . . . 12

5.3.2 Odds-Ratio . . . 12

5.4 Comparaison des performances CART, Logit, RF . . . 13

6 Conclusion **14**

1 Introduction

Ce jeu de données a été réalisé par Barry Becker à partir de la base de données du recensement de 1994. Il porte sur 32561 américains et vise à prédire, à partir d'une dizaine d'attributs, si un individu gagne plus de cinquante mille dollars par an ou non.

2 Le jeu de données

Le jeu de données possède 15 variables :

age	workclass	fnlwgt	education	education_num	marital.status	occupation
quantitatif	qualitatif	supprimé	qualitatif	supprimé	qualitatif	qualitatif

relationship	race	sex	capital.gain	capital.loss	hoursperweek	nativecountry	in
qualitatif	qualitatif	qualitatif	quantitatif	quantitatif	quantitatif	qualitatif	

Cependant, on ne va garder que 13 variables. En effet, on enlève la variable "education_num", car elle est récurrente avec la variable "education" ; ainsi que la variable "fnlwgt" car elle représente le nombre de personnes que l'auteur du recensement considère comme étant représentés par l'individu en question.

Donc finalement, on a 13 variables dont seulement 4 sont quantitatives.

3 Réduction de dimensions

3.1 ACM

Nous allons dans un premier temps faire une ACM sur les 9 variables qualitatives. Nous allons donc dans un premier temps ignorer les 4 autres variables quantitatives.

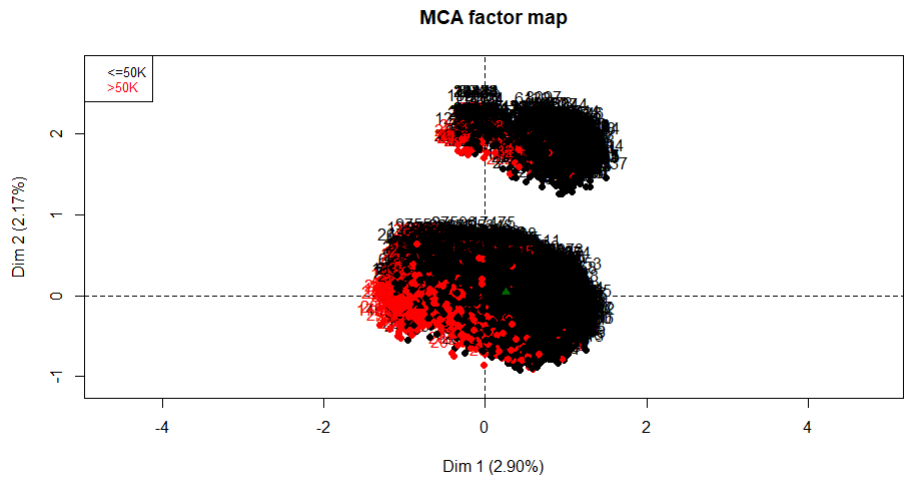

FIGURE 1 – ACM-Représentation des individus

On constate que le jeu de données ne se prête pas très bien à une ACM comme en témoigne la faible proportion d'inertie expliquée par les 2 premières composantes principales

(de l'ordre de 2%). Il est dès lors très compliqué d'intérpréter les résultats. On remarque dans cette nouvelle base que le jeu de donné est divisé en deux groupes. Cette division cependant, ne correspond pas à la séparation entre les deux modalités qui nous intéressent.

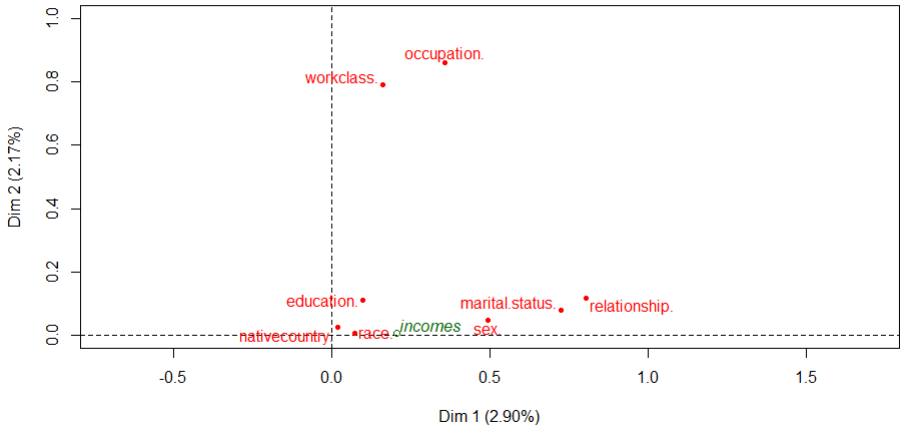

FIGURE 2 – ACM-Représentation des variables

Pour ce qui est de la représentation des variables, on retrouve une certaine cohérence : Comme on pourrait s'y attendre, certaines variables telles que "catégorie professionnelle" (workclass) et "métier" (occupation) ou encore "situation familiale" (marital-status : marrié, divorcé...) et "relations familiale" (a un enfant, célibataire...) sont corrélées.

3.2 FAMD

On utilise maintenant la fonction FAMD pour réduire l'ensemble des variables (quantitatives et qualitatives) à 2 composantes principales. L'avantage de cette méthode est qu'elle nous permet de prendre en compte toutes nos variables en même temps.

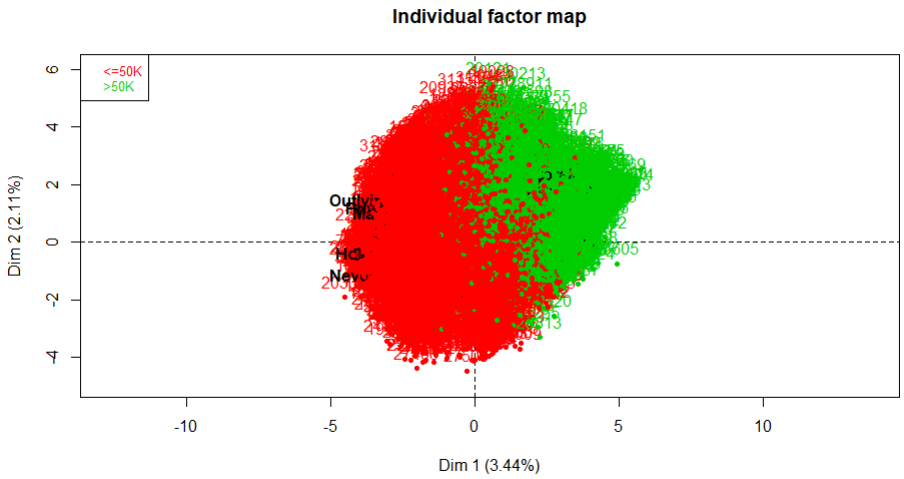

FIGURE 3 – ACM-Représentation des individus

On remarque tout d'abord que cette fois-ci, la représentation des individus dans la base des 2 premières composantes principales fait bien apparaître la distinction en les 2 modalités qui nous intéressent. On déduit ainsi que le fait de gagner plus de 50 milles dollars par an est en quelque sorte corrélé positivement à la première composante principale. Plus précisément, l'axe induisant la richesse semble être une combinaison linéaire de coefficients positifs, entre les deux axes principaux (davantage le premier que le deuxième).
Concernant les variables, on retrouve quelques corrélations entre variables telles "relationship" et "marital-status", mais on se rend compte que l'inertie a été répartie différemment puisque nous n'avons plus la proximité entre par exemple "workclass" et "occupation". Cela peut sans doute s'expliquer en partie par le fait que l'on prend cette fois également en compte les 4 variables quantitatives.

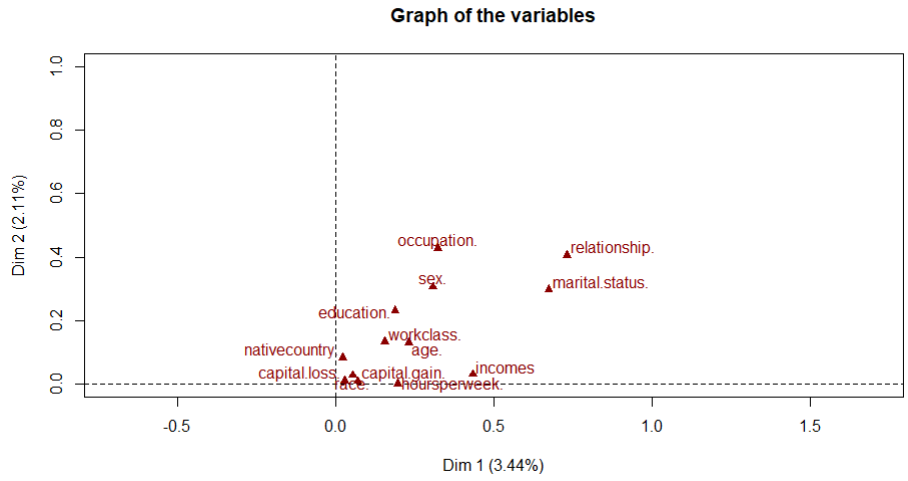

FIGURE 4 – ACM-Représentation des variables

Mais il faut toutefois prendre ces résultats avec des pincette puisqu'une fois encore, les 2 premières composantes principales n'expliquent qu'une faible partie de l'inertie (environ 5% les 2 réunies).
De plus, comme le montre le graphique 5, les variables quantitatives sont très mal représentées dans la nouvelle base puisqu'elles sont clairement à l'intérieur du cercle de corrélation.

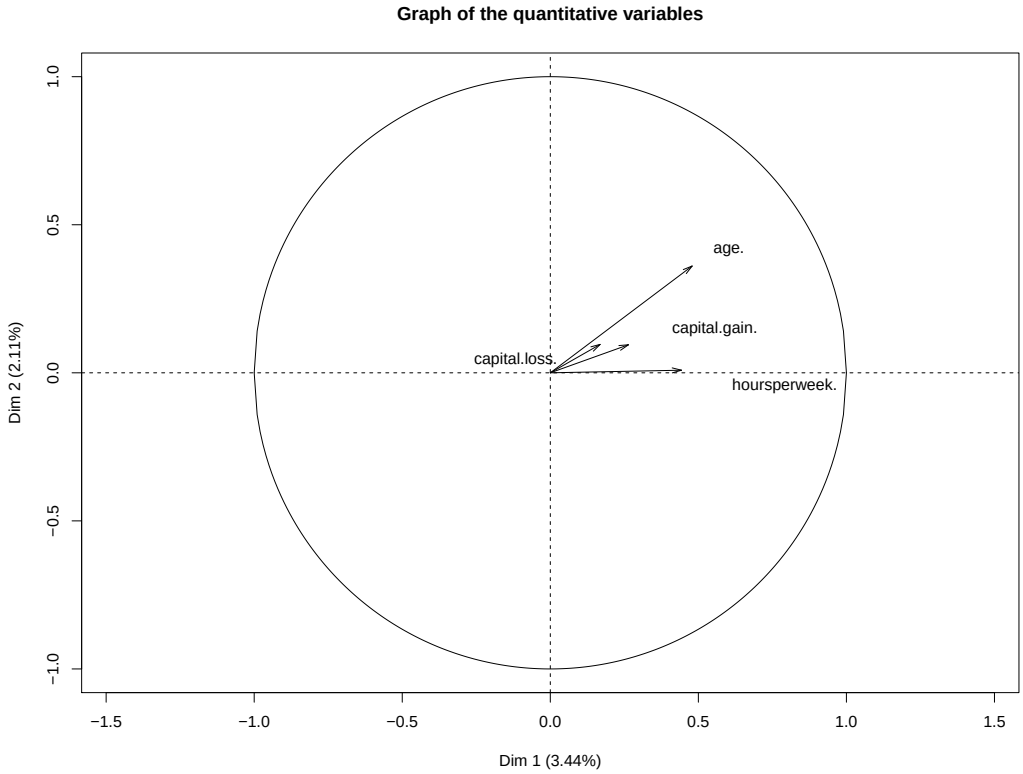

FIGURE 5 – Cercle de correlation FAMD

4 Apprentissage non-supervisé

Nous avons des données mixtes. Nous ne pouvons donc pas directement appliquer les méthodes de classification supervisée vues en cours. Nous allons donc utiliser les 2 premières composantes principales obtenues à la section précédente (ACM et FAMD).

4.1 K-means sur 2 classes

Nous allons dans un premier temps utiliser la méthode des K-means sur 2 classes. Pour les 2 premières composantes de l'ACM, on obtient la figure suivante :

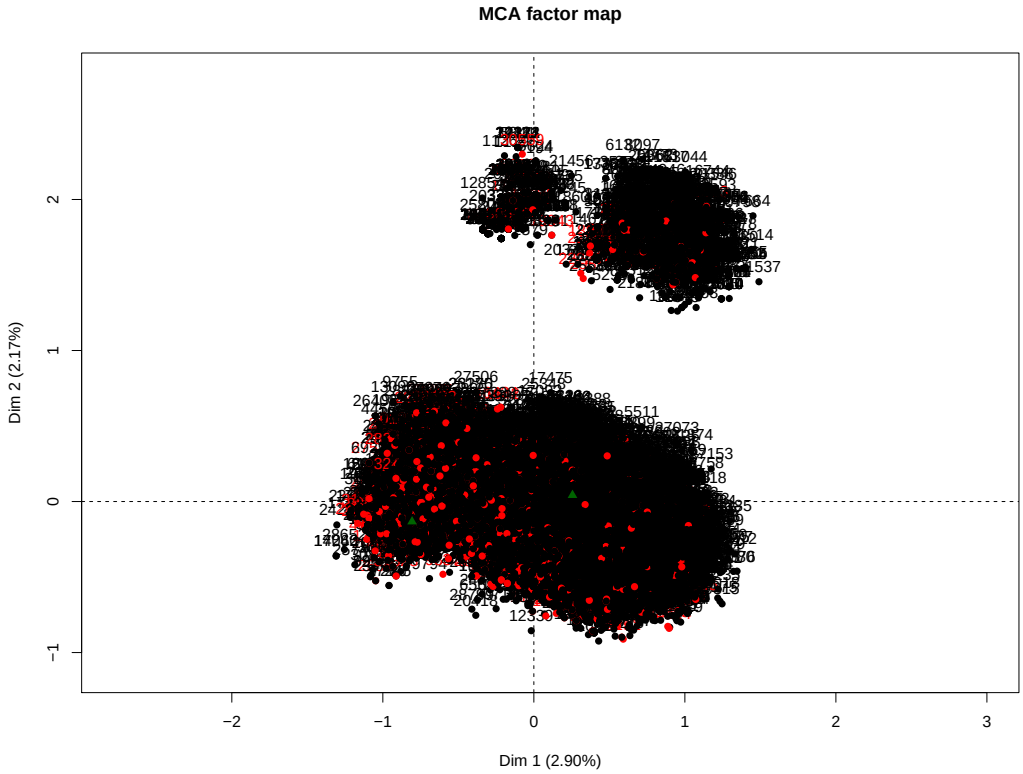

FIGURE 6 – Classification Kmeans : rouge classe 1 : noir ; classe 2 : rouge

Nous voulons vérifier si le clustering obtenu coïncide avec la séparation entre les deux modalités de la variable "incomes" :

Kmeans \ Incomes	<=50k	>50k
1	23974	6712
2	746	1129

FIGURE 7 – table de comparaison cluster-modalités à partir de l'ACM

Kmeans \ Incomes	<=50k	>50k
1	16969	568
2	7751	7273

FIGURE 8 – table de comparaison cluster-modalités à partir de l'FAMD

On constate que quelque soit la méthode de réduction de données que l'on a utilisée, les 2 groupes obtenus par Kmeans ont une tendance à séparer les individus en fonction de leur revenu. Dans les 2 cas cependant, cette séparation reste trop grossière. Ainsi le premier groupe obtenu par le kmeans basé sur les composantes principales de l'ACM et des variables quantitatives ; censé correspondre aux individus gagnant moins de 50 milles dollars par an ; possède plus d'individus ayant l'autre modalité que le cluster censé correspondre aux individus qui gagnent plus de 50 milles dollars par an. On retrouve un problème similaire pour ce qui est de la FAMD.

4.2 K-means sur plusieurs classes.

On va essayer de trouver un nombre de classes K tel que l'inertie intra-classe soit suffisamment faible. Pour ce faire on lance la méthode pour des valeurs de K allant de 1 à 10.

A chaque fois, on lance 100 initialisations aléatoires et on garde la meilleure. En traçant la proportion d'inertie intra-classe par rapport à l'inertie totale en fonction du nombre de groupes K, on obtient :

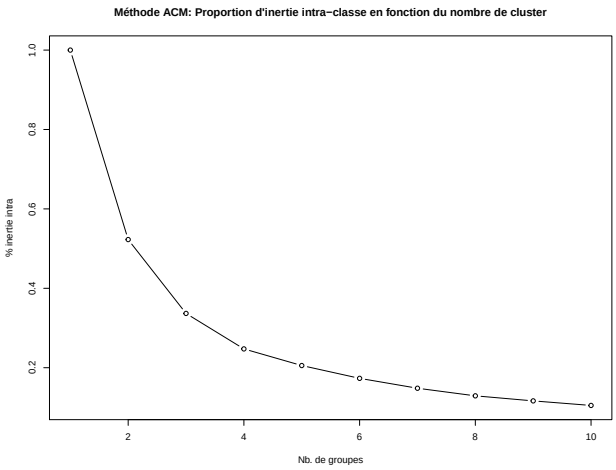

FIGURE 9 – inertie intra-classe / inertie totale. A partir de l'ACM

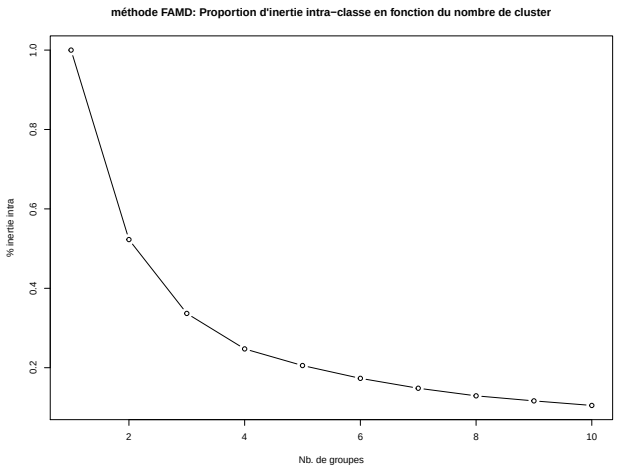

FIGURE 10 – inertie intra-classe / inertie totale. A partir du FAMD

Au-delà de 4 ou 5 groupes, ce rapport est inférieur à 0.3. Cela est vérifié quelle que soit l'origine des composantes principales qui on servit à la réalisation du Kmean. On peut donc supposer qu'une classification en 5 classes est une bonne clusterisation.

4.3 CAH

4.3.1 Résultat et intérprètation

Nous allons à présent appliquer la méthode du CAH en utilisant la fonction "HCPC" du package "FactoMiner". Cependant, notre jeu de donnés est trop gros en terme de place mémoire pour cette fonction. Nous avons donc tronqué notre jeu de donnés à 7000 individus. On n'utilise désormais que les composantes principales obtenus via FAMD. Nous obtenons le dendogramme suivant :

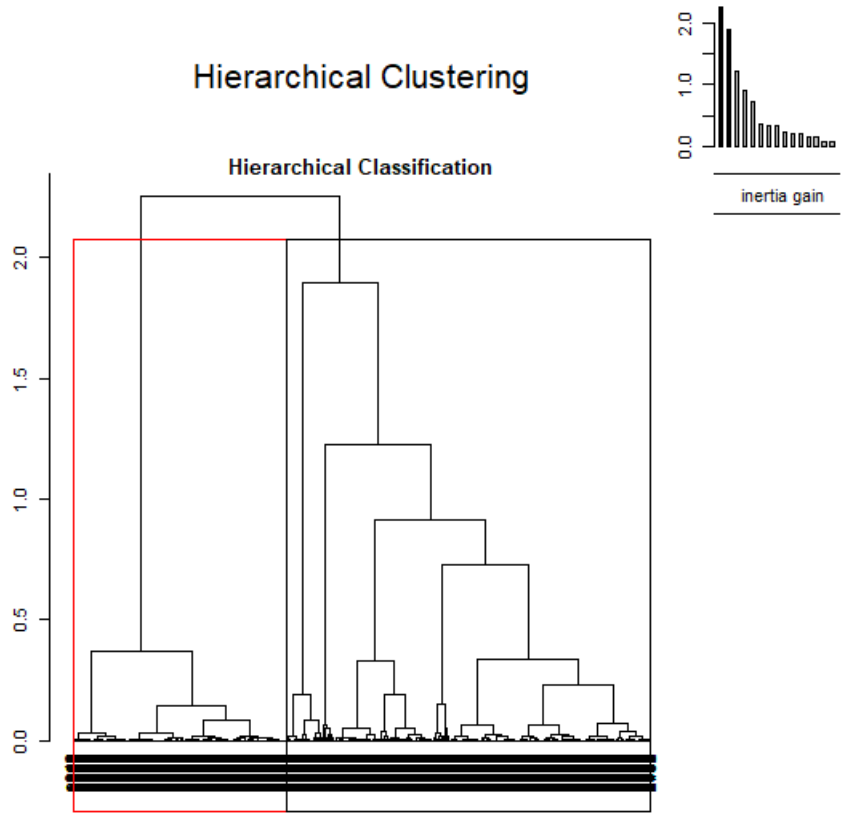

FIGURE 11 – Dendogramme CAH

Comme on l'a fait avec la méthode des K-means, on va tester si le clustering de la méthode CAH sépare bien le jeu de donnés en fonction des deux modalités qui nous interéssent :

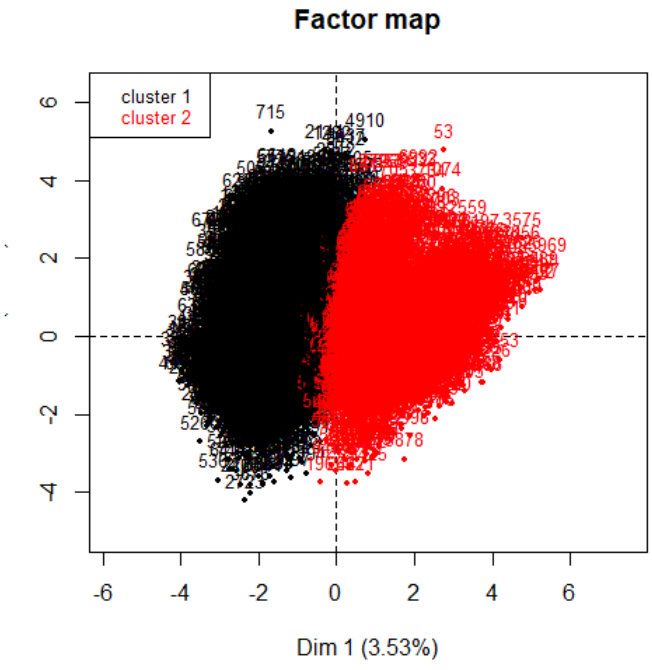

FIGURE 12 – CAH

on constate qu'à première vue, les clusters trouvés par la méthode du CAH coincident à peu près avec les 2 modalités d'intérêt. En effet, si on compare ce graphique avec celui que nous avons eu en représentant les individus sur les axes principaux obtenus par la FAMD (figure 3) ; on constate que sur l'emplacement des individus appartenant à la classe "<=50K" coïncident globalement avec les individus placés dans le cluster 1 (côté "gauche" du graphique) ; tandis que ceux de la classe ">50K" coïncident à peu près avec ceux classés dans le cluster 2 (côté "droit" du graphique).

4.3.2 Comparaison avec le K-means

Pour comparer les 2 méthodes de manière pertinente, nous avons relancé la méthode des K-means sur les mêmes données que la CAH ; c'est-à-dire les 7000 premiers individus. Si on compare les clusters obtenus avec les 2 méthodes, on constate qu'ils sont très proches :

CAH \ kmeans	1	2
1	3732	39
2	31	3198

FIGURE 13 – Table de comparaison des clusterings CAH/K-means

Pour aller encore plus loin, on peut calculer les Rand index qui est un coefficient mesurant la similarité entre deux clusterings. Dans notre cas, on obtient un rand index de 0.96 ce qui indique une forte similarité entre les 2 clusterings.

5 Apprentissage supervisé

Nos variables étant mixtes, nous n'allons pas utiliser des méthodes telles que l'AFD, le LDA ou encore le QDA ; mais nous allons baser notre analyse avec les approches non-paramétriques et semi-paramétriques.

5.1 CART

5.1.1 Arbre de classification standard

Dans cette section, on va créer un arbre de classification via la méthode CART. On va dans un premier temps lancer la méthode sans se soucier de la hauteur et de la complexité de l'arbre :

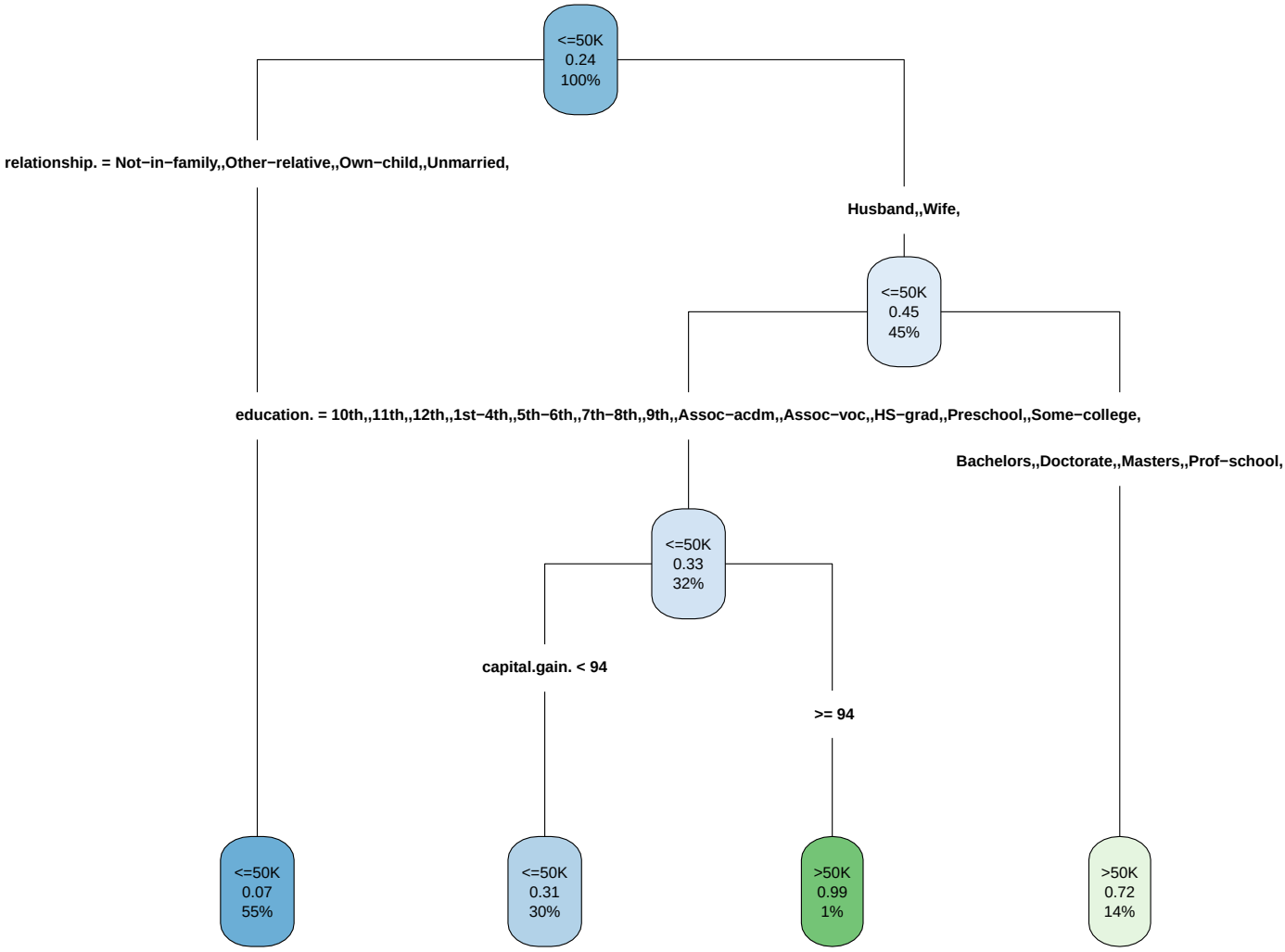

FIGURE 14 – Caption

On remarque que :

— Le premier facteur de distinction utilisé par la méthode est la situation familiale : Les individus étant dans une situation familiale moins "conventionnelle" (célibataire, ne vivant pas en famille...) sont selon cet arbre susceptible à 93% de gagner moins de 50 mille dollars par an. Cette situation concerne 55% de la population étudiée.

— Le second facteur est le niveau d'étude : Les personnes ayant fait des études longues (et étant mariés) ont 72% de chance de gagner plus de 50 mille dollars. Cela ne représente que 14% de la population étudiée.

— Ceux qui sont mariés et qui ont fait des études plus courtes sont séparés en fonction du montant des revenus issues d'un investissement ou d'une vente de propriété, ce qu'on appelle le capital gain. Les individus qui ont un "capital gain" supérieur ou égal à 94 milles dollars par ans ont 99% de chance de gagner plus de 50 milles dollars. Ils ne représentent cependant que 1% de la population étudiée.

5.1.2 Arbre optimal

Toutes les observations faites précédemment concernant l'arbre standard semblent logique ce qui nous conforte quant à la qualité de l'arbre réalisé. On pourrait toutefois affiner cet arbre en augmentant la pureté des feuilles notamment. Mais un arbre trop affiné devient vite illisible et inutilisable. Pour trouver un compromis, entre ces deux bords, on trace le graphique du taux d'erreur obtenu par validation croisée, en fonction du "cp" :

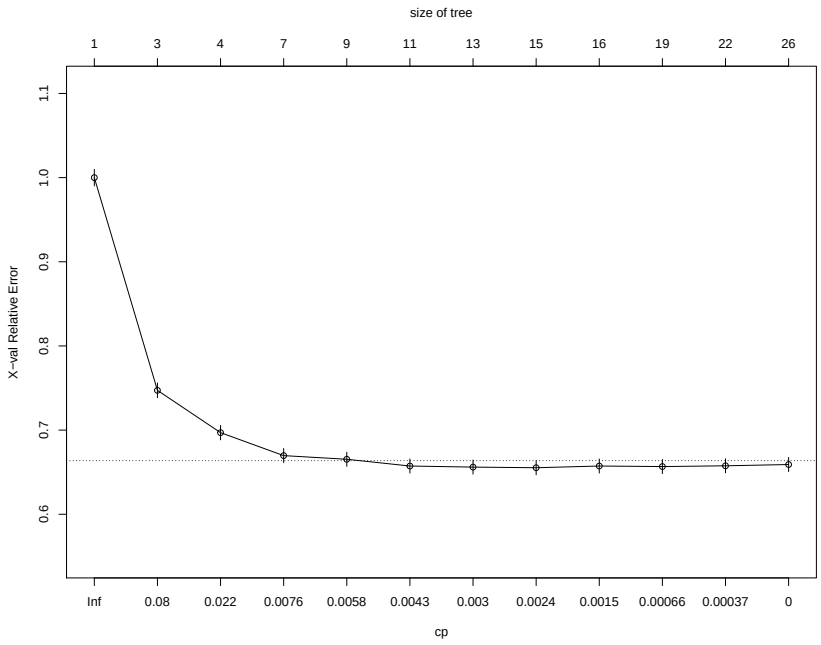

FIGURE 15 – erreur en fonction du CP

On constate que le taux d'erreur ne diminue plus à partir d'un cp inférieur à 0.0043. On crée donc un arbre optimal en prenant la valeur de cp que nous venons de trouver.

5.2 RANDOM FOREST

Pour parer la grande instabilité de la méthode CART quant aux données d'entraînement, on met en place un apprentissage par la méthode de la forêt aléatoire avec un nombre d'arbres par défaut égale à 500. On trace l'estimation de l'erreur de classification en fonction du nombre d'arbres :

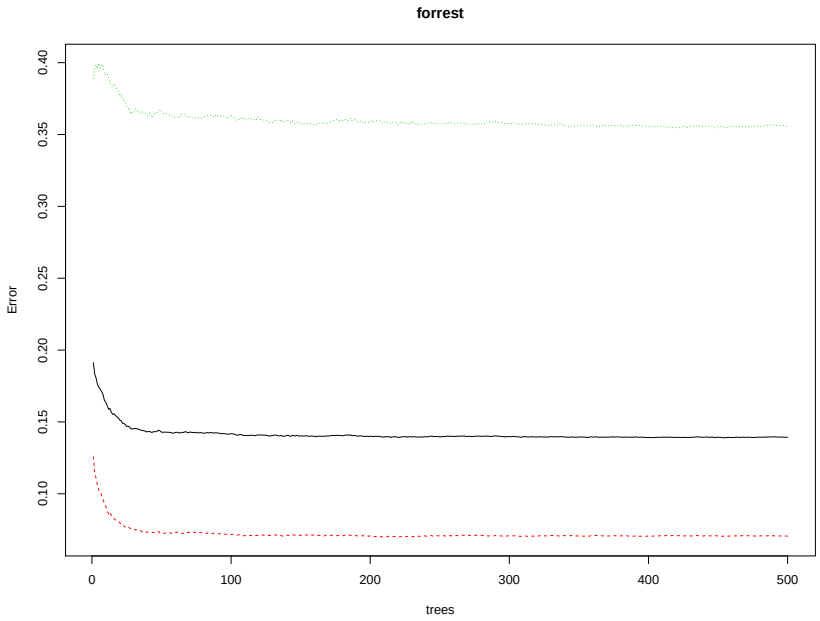

FIGURE 16 – Estimation de l'erreure de classification en fonction du nombre d'arbres

On constate que l'erreur de classification stagne à partir d'environ 200 arbres. On aurait donc pu prendre un nombre d'arbres plus faible. Cependant, l'erreur de classification des individus de la classe ">50K" demeure trop importante. On aurait pu augmenter le nombre de variables testées à chaque split.

5.3 Régression logistique

5.3.1 Tests

Nous mettons à présent en place un modèle de régression logistique pour expliquer et prédire la variable "Incomes". Pour s'assurer qu'il y a au moins une variable explicative significative, nous mettons en place le test de rapport de vraisemblance.
↪ Nous obtenons une p-value = $2.2e^{-16}$. Nous pouvons donc affirmer avec un risque de se tromper l'inférieur à 5% qu'il y a au moins une variable significative.
Cependant, si on regarde la contribution individuelle des variables en se référant à la p-value associée au test de Wald, on constate que toutes les modalités des variables qualitatives n'apportent pas de réel contribution. Nous avons donc voulu mettre en place la procédure "step" (avec direction=both) qui se base sur le critère AIC afin de sélectionner un bon modèle. Cette méthode n'a cependant pas trouvé de meilleur modèle de régression logistique et a renvoyé le même modèle que précédemment.

5.3.2 Odds-Ratio

Le test de Wald nous permet de voir la contribution individuelle des variables, mais il faudrait aussi être en mesure de comprendre comment elles contribuent. Pour ce faire, on peut utiliser l'Odds-ratio :

FIGURE 17 – OR_sex

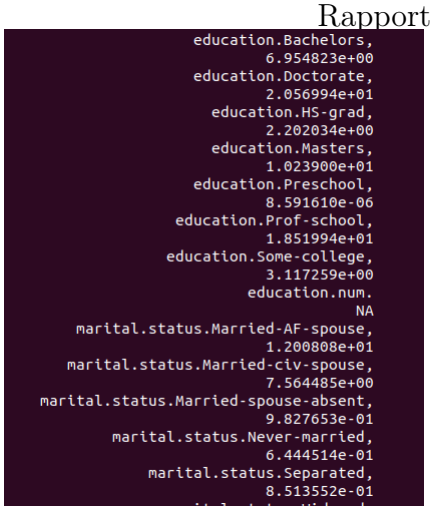

FIGURE 18 – OR_famille

On peut voir par le OR qu'en passant d'une modalité à une autre, les chances de gagner plus de 50k dollars fluctuent. Par exemple, selon ce modèle, être un homme donne 2.37 fois plus de chance de gagner plus de 50 mille dollars par ans que si l'on est une femme. De même être marié augmente 12 fois les chances d'appartenir à la catégorie ">50K" tandis qu'être doctorant les augmente de 20. On retrouve ici les conclusions qu'on a faites avec l'arbre de classification de la méthode CART.

5.4 Comparaison des performances CART, Logit, RF

La source où nous avons trouvé nos données d'apprentissage mettait également à disposition un jeu de donné test de 16281 individus. C'est donc ce jeu de donné que nous avons utilisé pour la phase de prédiction.

On va comparer les performances des trois méthodes de classification supervisée que nous avons essayé. Pour ce faire, on utilise les trois modèles apprit pour prédire la modalité de la variable "incomes" des individus du jeu de donné test. On obtient les matrices de confusions suivantes :

classcart	<=50k	>50k
<=50K	11511	1613
>50K	924	2233

FIGURE 19 – cart matrice de confusion

forestclass	<=50k	>50k
<=50K	11545	1500
>50K	890	2346

FIGURE 20 – forest matrice de confusion

classlogit	<=50k	>50k
0	11538	1610
1	897	2236

FIGURE 21 – logit matrice de confusion

On peut donc calculer la précision de ces méthodes :

— CART= 0.844
— Random Forest=0.853
— Regression logistique=0.846

On trace la courbe ROC :

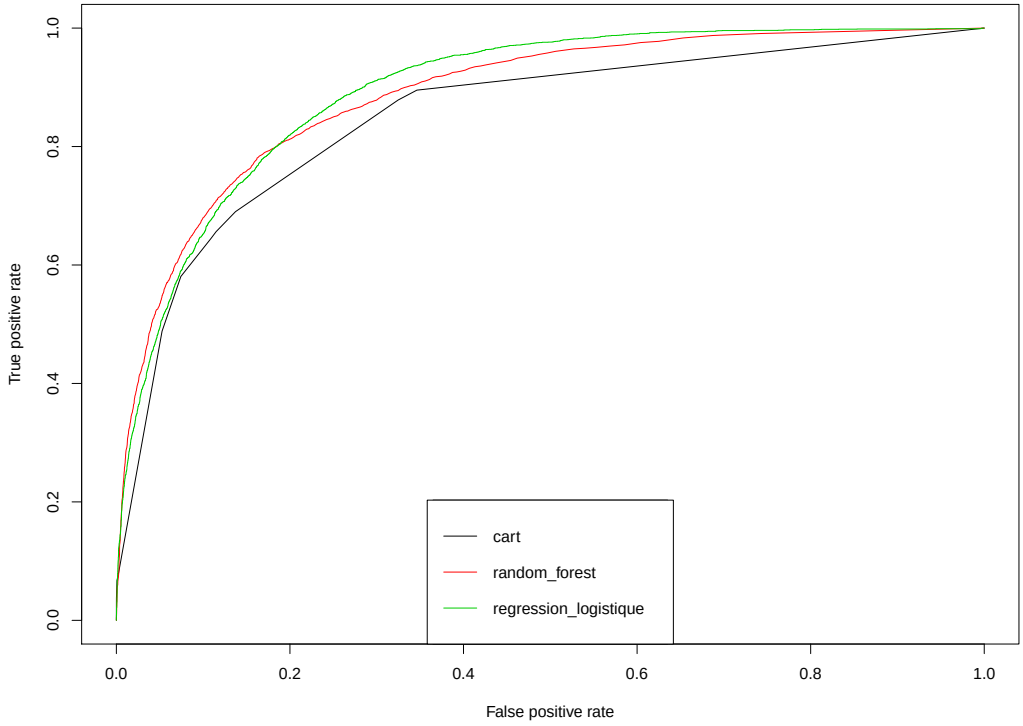

FIGURE 22 – ROC

On peut voir que la régression logistique est légèrement mieux adaptée que cart et random forest pour un taux de faux-positifs supérieur à 0.2 . (fig 22)

6 Conclusion

Pour conclure, selon notre étude, il semblerait que les facteurs importants pour gagner plus de 50 milles dollars (aux Etat-Unis en 1994) soient principalement la situation familliale et le niveau d'études. Le sexe et les revenus annexes (investissement, vente de propriété...) jouent un rôle non-négligeable également.

Notre jeu de donnés est un peu déséquilibré (environ trois fois plus d'individus de la classe "<=50K" que d'individus ">50K"). Une piste d'amélioration pourrait être de ré appliquer toutes ces méthodes à un jeu de donné plus équilibré.