Rapport TP1

Thibault Cart & Rami Albadri

Table des matières

Introduction

Le problème de ce TP est qu'on ne sait pas quels sont les facteurs les plus importants qui déterminent si une personne gagne plus ou moins de 50K$

Analyse préliminaire

Il y a 15 variables La variable cible est « income » car c'est ce qu'on recherche et elle est qualitative puisque 2 possibilités, soit plus que 50k ou moins que 50k

Attribute	**Type**	**Justification**
age	Quantitative	Valeur numérique continue (âge), adaptée aux calculs statistiques.
workclass	Qualitatif	Catégories d'emploi (p.ex., privé, fonction publique d'État, etc.)
fnlwgt	Quantitative	Score de pondération démographique.
education	Qualitatif	Niveaux d'éducation non ordonnés (p.ex., licence, etc.)
education-num	Quantitative	Version numérique ordonnée du niveau d'éducation.
marital-status	Qualitatif	Catégories d'état civil (marié, divorcé, etc.).
occupation	Qualitatif	Type de profession (vente, support technique, etc.).
relationship	Qualitatif	Rôle au sein du foyer (époux, hors famille, etc.).
race	Qualitatif	Groupe ethnique (blanc, noir, etc.).
sex	Qualitatif	Sexe biologique (homme, femme).
capital-gain	Quantitative	Gains en capital en valeur numérique.
capital-loss	Quantitative	Pertes en capital en valeur numérique.
hours-per-week	Quantitative	Nombre d'heures travaillées par semaine.
native-country	Qualitatif	Country of origin (United-States, India, etc.).
income	Qualitatif	Classe cible (binaire : <=50k ou >50k

Certaines variables doivent-elles être supprimées ?

Non toutes valeurs qualitatives sont utiles

Y a-t-il des données manquantes ?

Nous avons supprimé toutes les lignes ayant une ou plusieurs champs qui était égale à « ? » afin d'avoir plus de cohérence dans les data.

```
Lignes initiales : 32561
Lignes supprimées car contenant un ou plusieurs ? : 2399
Lignes restantes : 30162
```

Analyse exploratoire TP1_A

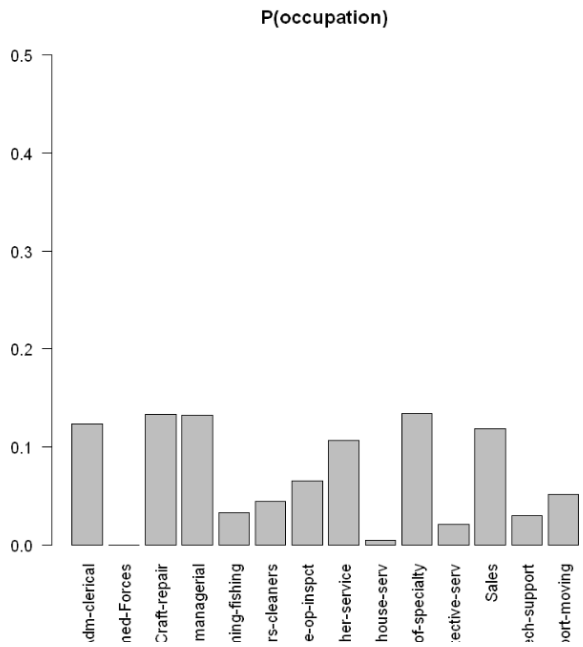
Attribue qualitatifs les plus utiles

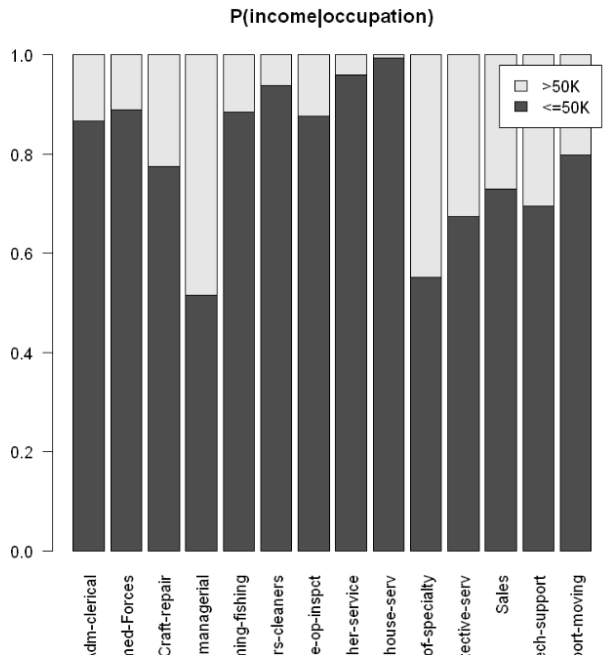
Nous avons sélectionné ces trois attribues car ils répondent au critère suivant.

Un nombre de données suffisants et reparties plus ou moins uniformément

Occupation

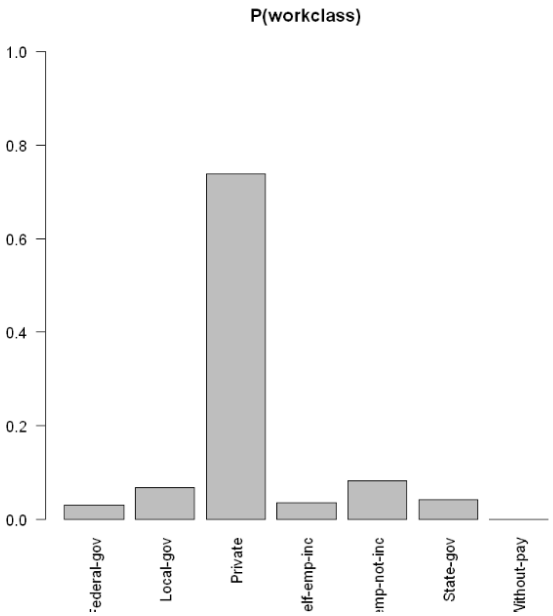
Nous pouvons voir que chaque occupation est liée à un revenu différent et de plus les données sont bien reparties entre chaque valeur et de façons plus ou moins homogène

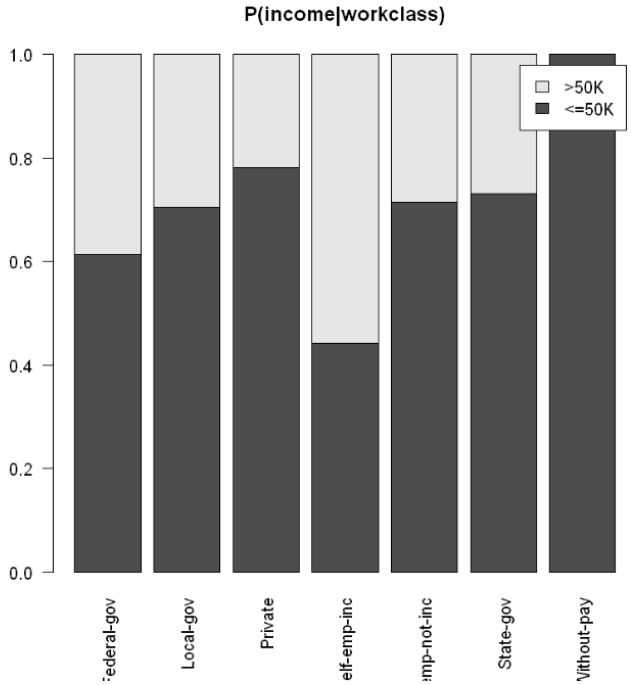




WorkClass

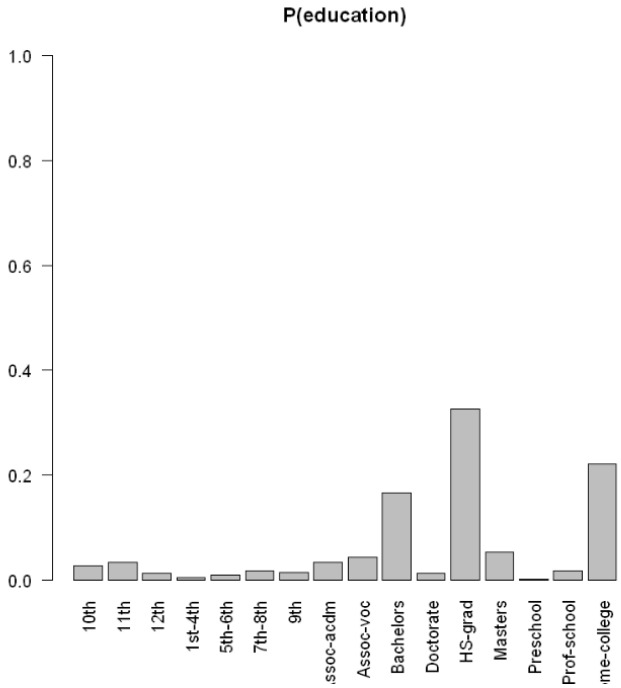
Les données sont moins reparties entre les différentes valeurs mais nous avons des résultats qui semble logique et peut influencé par la répartition

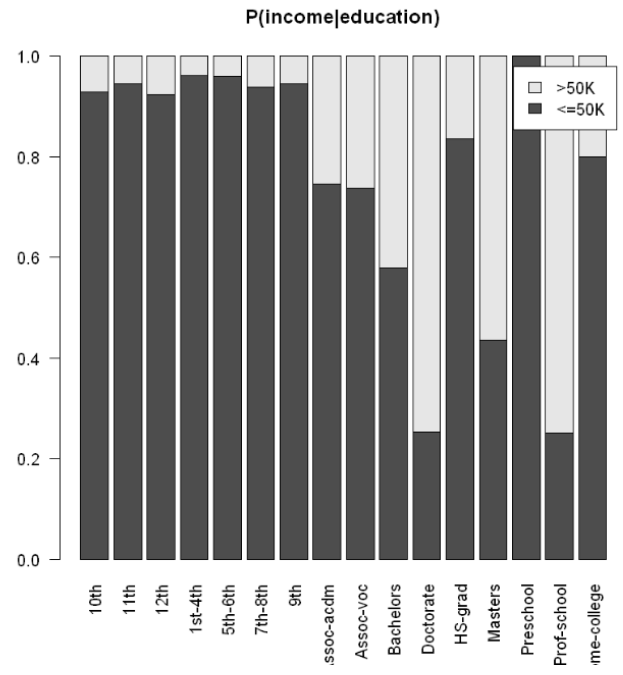




Education

Nous pouvons voir une distinction et une évolution net entre les personnes ayant poursuivit leurs études et ceux donc la scolarité c'est arrêté après l'High school.

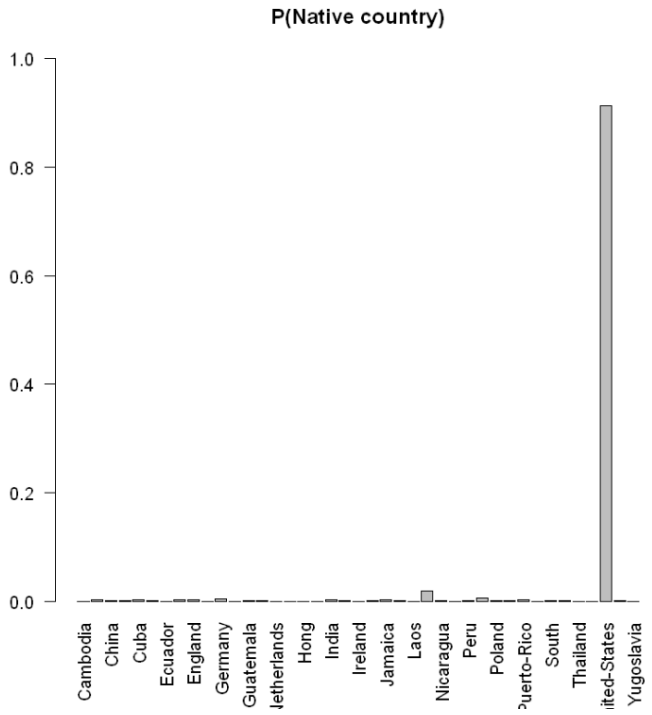


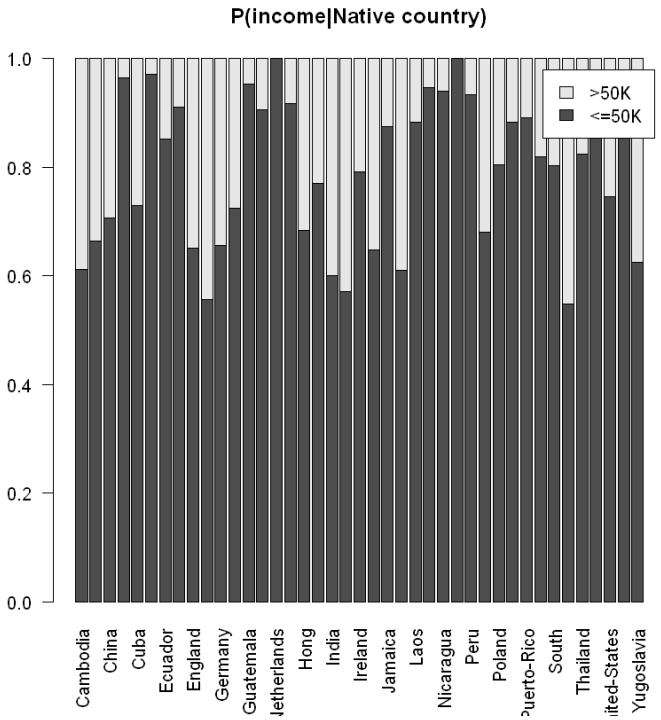


Attribue qualitatif peu impactant

Native Country

La répartition des données est vraiment mauvaise avec plus de 26'000 entré pours des Américains et les autres qui dépassent rarement les 100. Cela donne des données peu précises étant donné le faible nombre de personne prise en compte.





Analyse de la distribution conjointe et application du théorème de Bayes

1. Distribution conjointe P(occupation, income)

Nous avons sélectionné deux variables qualitatives issues du jeu de données :

- occupation : la profession de l'individu (par exemple : Exec-managerial, Sales, Tech-support, etc.),
- income : le revenu de l'individu (deux catégories : <=50K ou >50K), qui est notre **variable cible**.

Nous avons d'abord calculé la **distribution conjointe P(occupation, income)**, qui donne la probabilité d'observer chaque combinaison des deux variables.

Par exemple, si l'on obtient P(occupation = "Exec-managerial", income = ">50K") = 0.05, cela signifie que 5 % des individus dans l'échantillon sont des cadres dirigeants gagnant plus de 50K.

2. Distributions marginales P(occupation) et P(income)

Les distributions marginales sont déduites de la distribution conjointe :

- **P(occupation)** est obtenue en **sommant les lignes** de la distribution conjointe (on ignore la variable income) :
- **P(income)** est obtenue en **sommant les colonnes** de la distribution conjointe (on ignore la variable occupation) :

Ces distributions donnent la fréquence globale des différentes modalités.
Par exemple, P(occupation = "Sales") = 0.12 signifie que 12 % des individus travaillent dans la vente.

3. Distributions conditionnelles P(occupation | income) et P(income | occupation)

À partir de la distribution conjointe et des distributions marginales, nous pouvons calculer :

P(occupation | income) : la probabilité qu'un individu exerce une certaine profession **étant donné** son revenu.

On divise chaque **colonne** de la table conjointe par la distribution marginale P(income) :

P(income | occupation) : la probabilité qu'un individu ait un certain revenu **étant donné** sa profession.

On divise chaque **ligne** de la table conjointe par P(occupation) :

4. Exemple du théorème de Bayes

Le **théorème de Bayes** permet de relier les deux probabilités conditionnelles calculées ci-dessus :

$$P(occupation|income) = \frac{P(income|occupation) \cdot P(occupation)}{P(income)}$$

Prenons un exemple fictif basé sur des données réalistes :

- P(income = ">50K" | occupation = "Exec-managerial") = 0.48
- P(occupation = "Exec-managerial") = 0.10
- P(income = ">50K") = 0.25

Alors :

$$P(occupation = "Exec-managerial" | income = " > 50K") = \frac{0.48 \times 0.10}{0.25} = \frac{0.048}{0.25} = 0.192$$

Cela signifie qu'environ **19,2 % des individus gagnant plus de 50K occupent un poste de cadre dirigeant**.
Ce raisonnement inversé est au cœur des méthodes probabilistes de classification comme le **naive Bayes classifier**.

Conclusion TP1-A

Grâce à cette analyse, nous avons pu :

- Quantifier les liens entre les variables occupation et income dans le dataset,
- Extraire des relations conditionnelles utiles pour la compréhension du comportement des groupes professionnels,
- Appliquer et interpréter le théorème de Bayes avec un exemple simple et parlant.

Cette démarche peut facilement être reproduite pour d'autres variables comme education, marital.status, ou race.

Analyse exploratoire TP1_B

Pour savoir si un attribut est utile ou non, on doit regarder son score, et son score est calculé en fonction de son utilité par rapport à quel point il nous permet de distinguer les classes de la variable cible. Par exemple à quel point âge nous permet de savoir si une personne gagne >50K ou <50K. Si toutes les personnes entre 20 et 70 ans gagne tous avec peu d'exception >50K alors l'attribut âge n'est pas utile car il ne nous permettra pas de distinguer les classes de income.

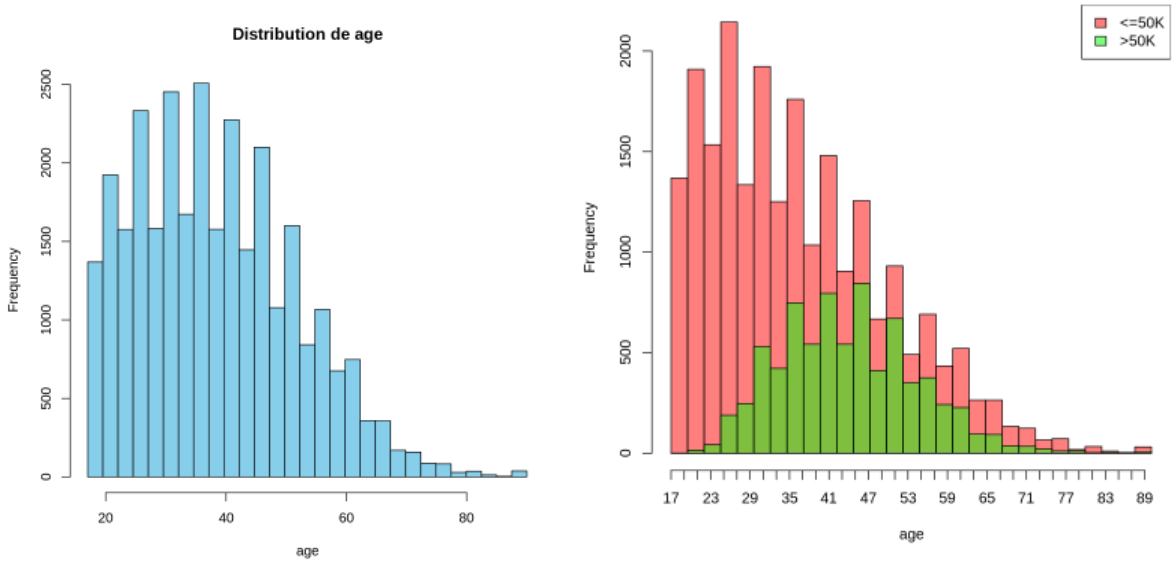
Sur l'image ci-dessous, on voit que « education.num » à le score le plus élevé avec 0.77 donc c'est le plus important, juste après vient « âge » avec le score de 0.55.

	Attribut	Moyenne	Variance	Moyenne_Classe1	Moyenne_Classe2
<=50K	age	38.4379	1.725137e+02	36.6081	43.9591
<=50K1	education.num	10.1213	6.502300e+00	9.6291	11.6064
<=50K2	capital.gain	1092.0079	5.485215e+07	148.8938	3937.6798
<=50K3	capital.loss	88.3725	1.634518e+05	53.4480	193.7507
<=50K4	hours.per.week	40.9312	1.435153e+02	39.3486	45.7066

	Variance_Classe1	Variance_Classe2	Score
<=50K	181.2883	1.054513e+02	0.5597
<=50K1	5.8252	5.608700e+00	0.7754
<=50K2	876791.7959	2.069312e+08	0.5116
<=50K3	96263.3866	3.513954e+05	0.3470
<=50K4	142.8147	1.152675e+02	0.5307

On commence donc par analyser chaque variable quantitative pour voir si elles permettent de distinguer les individues selon leur revenu (income) sauf income car variable cible et fnlwgt car c'est juste un poids statistique pour la population US.

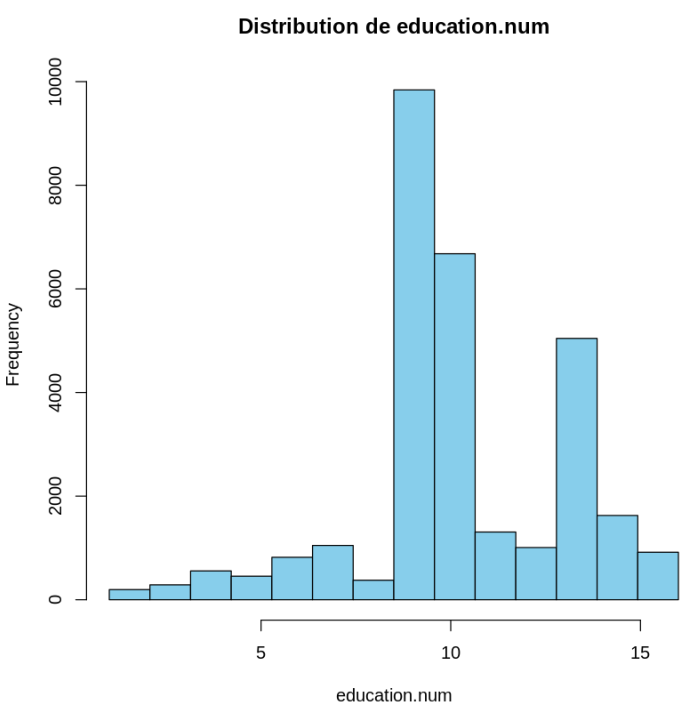
Dans les faits si on observe âge :

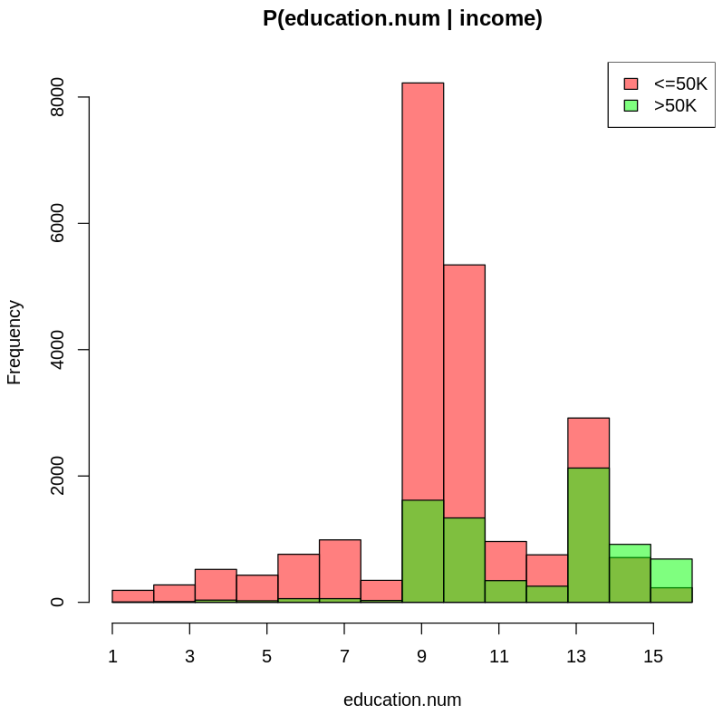


On voit dans le graphics de droit que les données ne sont pas parfaitement superposé, donc l'age est utiles pour faire de la classifications.

Prenons maintenant l'exemple de la variable education.num, qui représente le niveau d'éducation de manière numérique. On observe que la majorité des individus, qu'ils gagnent plus ou moins de 50K, ont un niveau d'éducation situé autour de 10. Toutefois, lorsqu'on compare les moyennes par classe, on remarque que ceux ayant un revenu supérieur à 50K ont en moyenne un niveau d'éducation plus élevé (environ 11.6 contre 9.6 pour ceux gagnant moins). De plus, il est important de souligner qu'il y a très peu, voire quasiment aucun individu ayant un niveau d'éducation inférieur à 9 qui perçoit un revenu supérieur à 50K. Cela montre que, même si la distribution générale est centrée autour de 10, un niveau d'éducation plus élevé reste fortement associé à une probabilité accrue d'avoir un salaire supérieur à 50K.

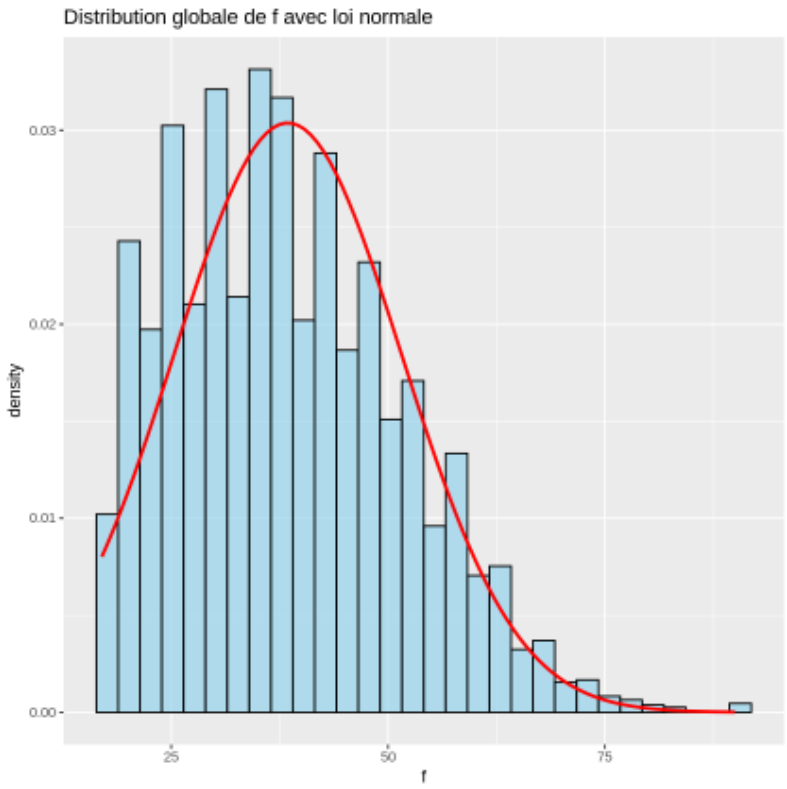
Cela suggère une forte corrélation entre un niveau d'éducation élevé et la probabilité de percevoir un revenu supérieur à 50K. Ce constat est appuyé par un **score d'importance de 0.7754**, faisant de education.num un des attributs les plus discriminants dans la classification du revenu.



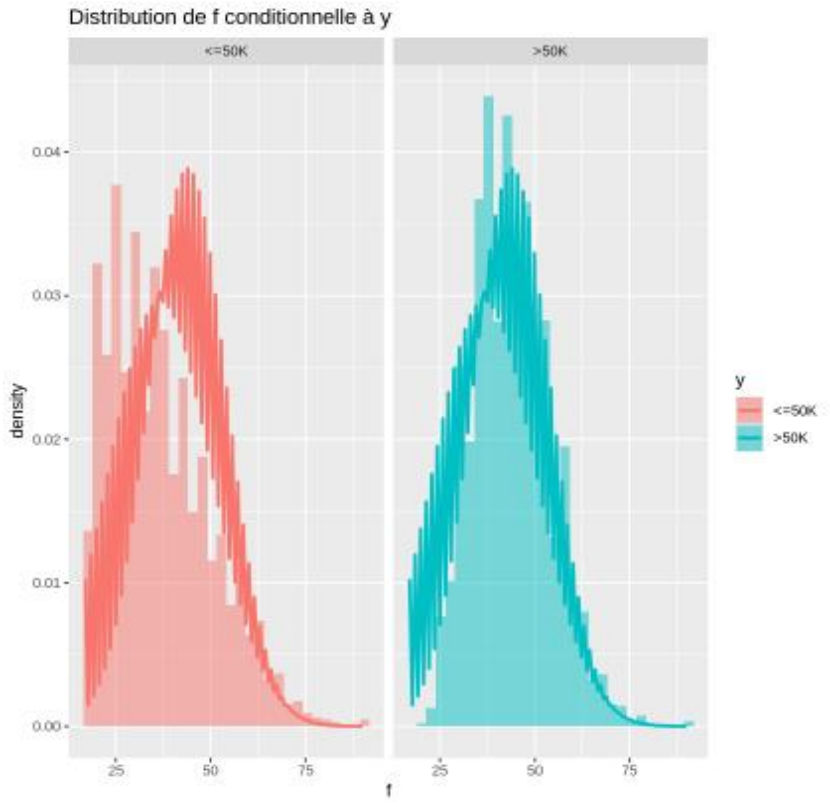


Grace a ce graphique scatter plot nous pouvons voir visuellement un lien entre l'âge, le niveau d'éducation ainsi que les revenus, dans ce graphique nous pouvons voir une poche de revenu supérieur à 50k à l'alentour des 55 ans et d'un niveau d'éducation d'environ 14-15, nous voyons aussi visuellement qu'il y'a très peu de personne gagnant plus de 50k en dessous du niveau d'éducation 10.

L’attribut âge a une distribution proche d’une loi normale, avec une moyenne d’environ 36 ans et un écart type de 13.



On observe que les distributions conditionnelles sont légèrement décalées mais restent de forme plus ou moins similaire. Donc cela suggère que l’âge suit approximativement une loi normale dans les deux courbes et surtout pour celle >50K.



Conclusion

À travers ces deux parties du TP, nous avons analysé en profondeur les attributs du dataset afin de comprendre quels facteurs influencent le plus le revenu d'un individu (<=50K ou >50K).

Dans la partie A, nous avons étudié les attributs qualitatifs. Nous avons vu que certains, comme occupation, education, ou workclass, présentent des liens clairs avec la variable cible. D'autres, comme native-country, sont peu exploitables car trop déséquilibrés. Nous avons également appris à utiliser les distributions conjointes, marginales et conditionnelles, et appliqué le théorème de Bayes pour raisonner de manière inversée (ex : P(profession | revenu)).

Dans la partie B, nous avons étudié les attributs quantitatifs, en utilisant des outils statistiques plus avancés. En calculant les scores d'importance pour chaque variable, nous avons pu identifier les deux plus discriminants : education.num et age. Grâce aux histogrammes et au scatter plot, nous avons pu visualiser clairement la séparation entre les classes, et comprendre comment ces variables contribuent à la prédiction. Enfin, nous avons confronté la distribution réelle d'un attribut (age) à une loi normale théorique.