# Rapport TP1

# Thibault Cart & Rami Albadri



# 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.

# Une image contenant texte, Police, capture d’écran, noir Le contenu généré par l’IA peut être incorrect.

# 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

Une image contenant texte, capture d’écran, Parallèle, typographie

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, diagramme, conception

Le contenu généré par l’IA peut être incorrect.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

Une image contenant texte, capture d’écran, diagramme, conception

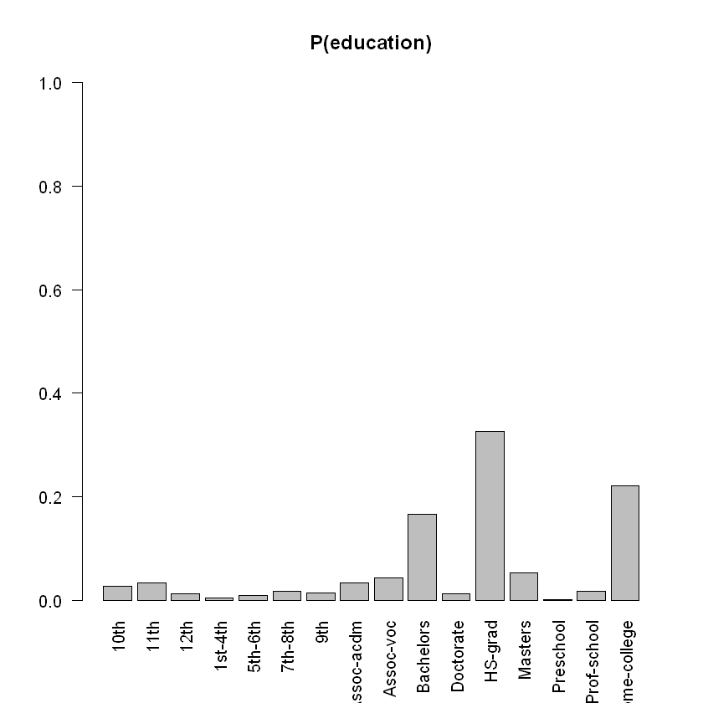
Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran

Le contenu généré par l’IA peut être incorrect.Les données sont moins reparties entre les différentes valeurs mais nous avons des résultats qui semble logique et peut influencé par la répartition

### Education

Une image contenant texte, capture d’écran, musique

Le contenu généré par l’IA peut être incorrect.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

Une image contenant texte, capture d’écran

Le contenu généré par l’IA peut être incorrect.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.

Une image contenant texte, capture d’écran, diagramme, Police

Le contenu généré par l’IA peut être incorrect.

## 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 :

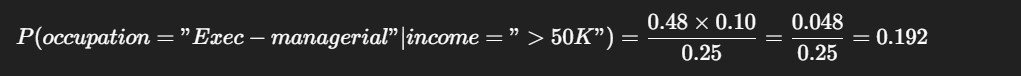
Une image contenant Police, texte, ligne, capture d’écran

Le contenu généré par l’IA peut être incorrect.

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 :

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**

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.