# 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

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.