Python for data analysis

Yassin HASSAN

Base de données **Census income**

Contexte de la base de donnée

- Pour ce projet, nous allons utilizer la base de données Census Income disponible sur le repertoire de l' UCIrvine Machine Learning Repository
- Le "US Adult Census" est un jeu de données de 48 842 entrées extraites de la base de donnée USCensus datant de 1994.

Cible à prédire

- Notre but est de prédire si les revenus d'un individu seront supérieur à 50 000 dollars par an sur la base de plusieurs caractéristiques socioculturelles continues dans le jeu de données.
- On cherche à prédire la colonne **income**.

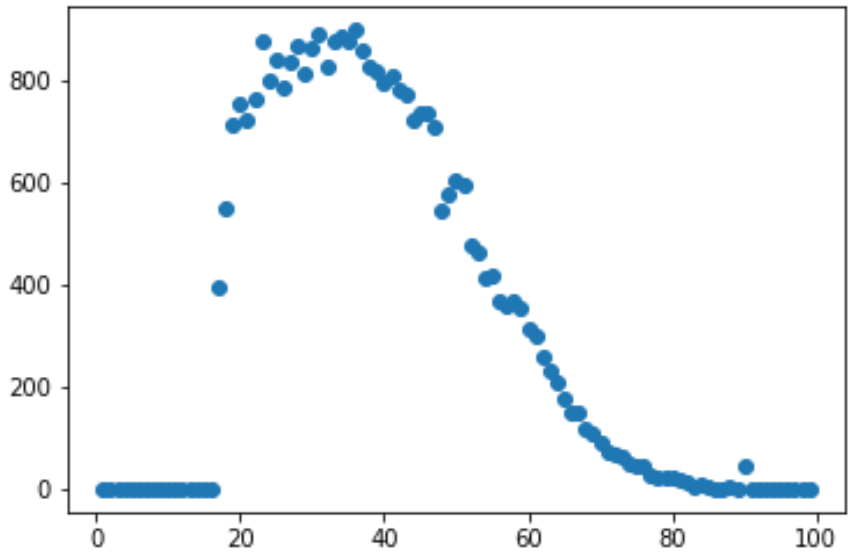
Analyse descriptive des données

- age – âge de l'individu
  - Nombre entier supérieur à 0
- type_employer – Type d'employeur de l'individu
  - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt – final weight, nombre de personnes le census estime que cette observation représente
  - Nombre entire supérieur à 0
- education – Plus haut niveau d'éducation de l'individu
  - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc,9th, 7th8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education_num – Plus haut niveau d'éducation de l'individu sous forme numérique
  - Nombre entire supérieur à 0
- marital – Etat civil de l'individu
  - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation – Occupation de l'individu
  - Tech-support, Craf-trepair, Other-service, Sales, Exec-managerial,Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical,Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv,Armed-Forces.

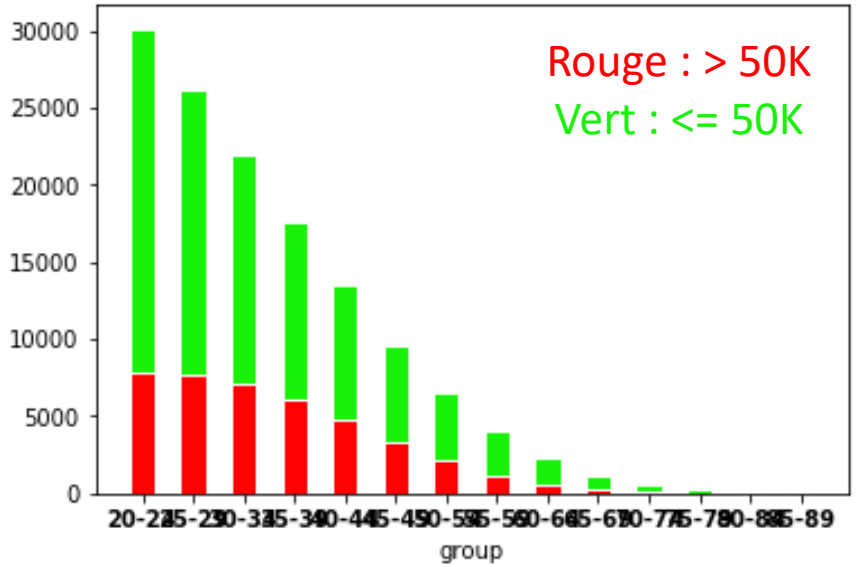
Analyse descriptive des données

- relationship – Observations réalisées des relations de l'individu
  - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- race – description ethnique de l'individu
  - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex – sexe de l'individu
  - Male, Female
- capital_gain – gains de capitaux enregistrés
  - Nombre entier supérieur ou égal à 0
- capital_loss – perte de capitaux enregistrés
  - Nombre entier supérieur ou égal à 0
- hr_per_week – heures de travail par semaine
  - Nombre entier supérieur ou égal à 0
- country – pays d'origine de l'individu
  - United-States, Cambodia, England, PuertoRico, Canada, Germany,Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran,Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal,Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia,Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador,Trinadad&Tobago, Peru, Hong, Holand-Netherlands
- income – valeur booleen indiquant si l'individu gagne plus de 50000 dollars par an
  - <=50k, >50k

Visualisation des données



Graphe représentant les âges des individus



Graphe représentant la répartition des revenus par tranche d'âge

Features disponibles

- Téléchargement des données (webscrapping)
- Visualisation des données
- Data-préparation des données
- Modélisation
- Optimisation des hyperparamètres
- Visualisation des performances

Récupération des données

La récupération des données peut être résumée en 3 étapes :

1. Récupération du contenu de la page web
2. Enregistrement local en supprimant les caractères inutiles
3. Transformation du fichier obtenu en dataframe

1

```
page = requests.get("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data")
#page.text
```

2

```
file = open('adult_data.csv', 'w')
text = page.text
text = text.replace(".", "")
text = text.replace(" ", "")
file.write(text)
file.close()
```

3

```
df = pd.read_csv("./adult_data.csv",delimiter=',', names=["age", "type_employer", "fnlwgt", "education",
                "education_num","marital", "occupation", "relationship", "race","sex",
                "capital_gain", "capital_loss", "hr_per_week","country", "income"])

df.head()
```