Base de Données : Adult (Census Income)

1. Identification générale

Nom : Adult (Census Income)
Source : UCI Machine Learning Repository
Donateur : Barry Becker (30 avril 1996)
Type : Données multivariées, mixtes (numériques et catégorielles)
Objectif : Prédire si le revenu annuel d'un individu dépasse 50 000 $/an.

2. Taille et caractéristiques des données

Nombre d'instances : 48 842
Nombre de variables : 14 (hors variable cible)
Valeurs manquantes : Oui ('?' dans certaines colonnes)
Format : CSV (séparé par des virgules)
Licence : Creative Commons BY 4.0

3. Description des variables

1. age : Âge de l'individu
2. workclass : Catégorie d'emploi
3. fnlwgt : Poids de l'échantillon
4. education : Niveau d'éducation
5. education-num : Niveau d'éducation (numérique)
6. marital-status : Statut marital
7. occupation : Type d'occupation
8. relationship : Statut de relation
9. race : Race
10. sex : Sexe
11. capital-gain : Gain en capital
12. capital-loss : Perte en capital
13. hours-per-week : Heures travaillées par semaine
14. native-country : Pays d'origine

Cible : income (>50K ou ≤50K)

4. Utilisation et contexte

Utilisation : Classification binaire, détection de biais (sexe, race)
Domaines : Apprentissage automatique, équité algorithmique, économie du travail
Limites : Données anciennes (recensement 1994), seuil arbitraire de 50 000 $, biais démographiques potentiels.

5. Recommandations d'analyse

• Analyse descriptive : distributions, corrélations, visualisations
• Nettoyage : gérer les valeurs manquantes ('?')
• Encodage : one-hot encoding des variables catégorielles
• Modélisation : régression logistique, arbres de décision, forêts aléatoires
• Interprétation : importance des variables, inégalités de revenus

• Communication : rédiger un rapport clair et structuré en français.