# Adult income dataset

A widely cited KNN dataset as a playground

**About Dataset**
An individual's annual income results from various factors. Intuitively, it is influenced by the individual's education level, age, gender, occupation, and etc.

This is a widely cited KNN dataset. I encountered it during my course, and I wish to share it here because it is a good starter example for data pre-processing and machine learning practices.

**Fields**
The dataset contains 16 columns

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, 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.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, 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.

    Target filed: Income
        -- The income is divide into two classes: <=50K and >50K
    Number of attributes: 14
        -- These are the demographics and other features to describe a person

We can explore the possibility in predicting income level based on the individual's personal information.

**Acknowledgements**
This dataset named "adult" is found in the UCI machine learning repository
http://www.cs.toronto.edu/~delve/data/adult/desc.html

The detailed description on the dataset can be found in the original UCI documentation
http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html

**The Adult dataset**

The information is a replica of the notes for the abalone dataset from the **UCI** repository.

**1. Title of Database: adult**

**2. Sources:**

(a) Original owners of database (name/phone/snail address/email address)

US Census Bureau.

(b) Donor of database (name/phone/snail address/email address)

Ronny Kohavi and Barry Becker,
Data Mining and Visualization
Silicon Graphics.
e-mail: ronnyk@sgi.com

(c) Date received (databases may change over time without name change!)

05/19/96

**3. Past Usage:**

(a) Complete reference of article where it was described/used

@inproceedings{kohavi-nbtree,
author={Ron Kohavi},
title={Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid},
booktitle={Proceedings of the Second International Conference on Knowledge Discovery and Data Mining},
year = 1996,
pages={to appear}}

(b) Indication of what attribute(s) were being predicted

Salary greater or less than 50,000.

(b) Indication of study's results (i.e. Is it a good domain to use?)

Hard domain with a nice number of records.
The following results obtained using MLC++ with default settings
for the algorithms mentioned below.

| | **Algorithm** | **Error** |
|---|---|---|
| 1 | C4.5 | 15.54 |
| 2 | C4.5-auto | 14.46 |
| 3 | C4.5-rules | 14.94 |
| 4 | Voted ID3 (0.6) | 15.64 |
| 5 | Voted ID3 (0.8) | 16.47 |
| 6 | T2 | 16.84 |
| 7 | 1R | 19.54 |

| | | |
|---|---|---|
| 8 | NBTree | 14.10 |
| 9 | CN2 | 16.00 |
| 10 | HOODG | 14.82 |
| 11 | FSS Naive Bayes | 14.05 |
| 12 | IDTM (Decision table) | 14.46 |
| 13 | Naive-Bayes | 16.12 |
| 14 | Nearest-neighbor (1) | 21.42 |
| 15 | Nearest-neighbor (3) | 20.35 |
| 16 | OC1 | 15.04 |
| 17 | Pebls | Crashed. Unknown why (bounds WERE increased) |

**4. Relevant Information Paragraph:**

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

**5. Number of Instances**

- 48842 instances, mix of continuous and discrete (train=32561, test=16281)
- 45222 if instances with unknown values are removed (train=30162, test=15060)
- Split into train-test using MLC++ GenCVFiles (2/3, 1/3 random).

**6. Number of Attributes**

6 continuous, 8 nominal attributes.

**7. Attribute Information:**

15. age: continuous.
16. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
17. fnlwgt: continuous.
18. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
19. education-num: continuous.
20. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
21. occupation: Tech-support, Craft-repair, 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.

22. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
23. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
24. sex: Female, Male.
25. capital-gain: continuous.
26. capital-loss: continuous.
27. hours-per-week: continuous.
28. native-country: United-States, Cambodia, England, Puerto-Rico, 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.

class: >50K, <=50K

**8. Missing Attribute Values:**

7% have missing values.

**9. Class Distribution:**

Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)
Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)

**10. Notes for Delve**

1. One prototask (income) has been defined, using attributes 1-13 as inputs and *income level* as a binary target.
2. Missing values - These are confined to attributes 2 (workclass), 7 (occupation) and 14 (native-country). The prototask only uses cases with no missing values.
3. The income prototask comes with two priors, differing according to if attribute 4 (education) is considered to be nominal or ordinal.