“Adult data set”, also known as "Census Income" dataset

This data was extracted from the census bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html

Donor: Ronny Kohavi and Barry Becker, Data Mining and Visualization, Silicon Graphics.

48842 instances, mix of continuous and discrete (train=32561, test=16281)

45222 if instances with unknown values are removed (train=30162, test=15060)

Duplicate or conflicting instances : 6

Class probabilities for whole data set:

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

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

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))

Prediction task is to determine whether a person makes over 50K a year.

First cited in:

@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 of Knowledge Discovery and Data Mining}, year = 1996, pages={to appear}}

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Values** | **Description** |
| age | Numeric | Min: 17 Max: 90 | Age of the person |
| workclass | Nominal | Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay | Class of work |
| fnlwgt | Numeric | Min: 12285 Max: 1490400 | Final weight of how much of the population it represents |
| education | Nominal | 10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college | Education level |
| education\_num | Numeric | Min: 1 Max: 16 | Numeric education level |
| marital\_status | Nominal | Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed | Marital status  of the person |
| occupation | Nominal | Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving | Occupation  of the person |
| relationship | Nominal | Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife | Type of relationship |
| race | Nominal | Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White | Race of the person |
| sex | Nominal | Female, Male | Sex of the person |
| capital\_gain | Numeric | Min: 0 Max: 99999 | Capital gains obteined |
| capital\_loss | Numeric | Min: 0 Max: 4356 | Capital loss |
| hours\_per\_week | Numeric | Min: 1 Max: 99 | Average number of hour working per week |
| native\_country | Nominal | Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States, Vietnam, Yugoslavia | Country of origin |
| income | Nominal | <=50K, >50K | Income level |