UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

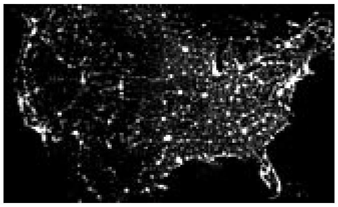
● Repository ○ Web Google™

View ALL Data Sets

Census-Income (KDD) Data Set

Download: Data Folder, Data Set Description

Abstract: This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau.



Data Set Characteristics:	Multivariate	**Number of Instances:**	299285	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	40	**Date Donated**	2000-03-07
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	139924

Source:

Original Owner:

U.S. Census Bureau
http://www.census.gov/
United States Department of Commerce

Donor:

Terran Lane and Ronny Kohavi
Data Mining and Visualization
Silicon Graphics.
terran '@' ecn.purdue.edu, ronnyk '@' sgi.com

Data Set Information:

This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables.

The instance weight indicates the number of people in the population that each record represents due to stratified sampling. To do real analysis and derive conclusions, this field must be used. This attribute should *not* be used in the classifiers.

One instance per line with comma delimited fields. There are 199523 instances in the data file and 99762 in the test file.

The data was split into train/test in approximately 2/3, 1/3 proportions using MineSet's MIndUtil mineset-to-mlc.

Attribute Information:

More information detailing the meaning of the attributes can be found in the Census Bureau's documentation To make use of the data descriptions at this site, the following mappings to the Census Bureau's internal database column names will be needed:

age AAGE
class of worker ACLSWKR
industry code ADTIND
occupation code ADTOCC
adjusted gross income AGI
education AHGA
wage per hour AHRSPAY
enrolled in edu inst last wk AHSCOL
marital status AMARITL
major industry code AMJIND
major occupation code AMJOCC
mace ARACE
hispanic Origin AREORGN
sex ASEX
member of a labor union AUNMEM
reason for unemployment AUNTYPE
full or part time employment stat AWKSTAT
capital gains CAPGAIN
capital losses CAPLOSS
divdends from stocks DIVVAL
federal income tax liability FEDTAX
tax filer status FILESTAT
region of previous residence GRINREG
state of previous residence GRINST
detailed household and family stat HHDFMX
detailed household summary in household HHDREL
instance weight MARSUPWT
migration code-change in msa MIGMTR1
migration code-change in reg MIGMTR3
migration code-move within reg MIGMTR4
live in this house 1 year ago MIGSAME
migration prev res in sunbelt MIGSUN
num persons worked for employer NOEMP
family members under 18 PARENT
total person earnings PEARNVAL
country of birth father PEFNTVTY
country of birth mother PEMNTVTY
country of birth self PENATVTY
citizenship PRCITSHP
total person income PTOTVAL
own business or self employed SEOTR
taxable income amount TAXINC
fill inc questionnaire for veteran's admin VETQVA
veterans benefits VETYN
weeks worked in year WKSWORK

Note that Incomes have been binned at the $50K level to present a binary classification problem, much like the original UCI/ADULT database. The goal field of this data, however, was drawn from the "total person income" field rather than the "adjusted gross income" and may, therefore, behave differently than the orginal ADULT goal field.

Relevant Papers:

N/A

Papers That Cite This Data Set[1]:

Eibe Frank and Geoffrey Holmes and Richard Kirkby and Mark A. Hall. Racing Committees for Large Datasets. Discovery Science. 2002. [View Context].

Stephen D. Bay. Multivariate Discretization for Set Mining. Knowl. Inf. Syst, 3. 2001. [View Context].

Nikunj C. Oza and Stuart J. Russell. Experimental comparisons of online and batch versions of bagging and boosting. KDD. 2001. [View Context].

Masahiro Terabe and Takashi Washio and Hiroshi Motoda. The Effect of Subsampling Rate on S 3 Bagging Performance. Mitsubishi Research Institute. [View Context].

Citation Request:

Please refer to the Machine Learning Repository's citation policy

[1] Papers were automatically harvested and associated with this data set, in collaboration with Rexa.info