**85.Census-Income (KDD)**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)

2. 数据库描述

【1.[数据集名称]数据集由[机构名或人名]采集；】The data used in our experiments were collected by E. Alpaydin, C. Kaynak, from Department of Computer Engineering,Bogazici University at July,1998.【2.用于[什么实验目的]】We used preprocessing programs made available by NIST to extract normalized bitmaps of handwritten digits from a preprinted form.【3】

【4】The database has 5620 samples, respectively belong to optdigits.tra with 3823 samples and optidigits.tes with 1797 samples. The categories of network system include seven categories, as shown in Table 1.

Table 1 Category Distribution of Network System [根据数据库绘制]

|  |  |  |  |
| --- | --- | --- | --- |
| Invasion Categories | optdigits.tra | optdigits.tes | Total Number of Samples |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Total number of samples in total |  |  | 299285 |

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

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

**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:   
  
  
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.

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 demographic and employment related variables.

## Data Characteristics

This data set contains weighted census data extracted from the 1994 and 1995 [Current Population Surveys](http://www.census.gov/cps/methodology/techdocs.html) 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.

More information detailing the meaning of the attributes can be found in the Census Bureau's [documentation](http://www.census.gov/cps/methodology/techdocs.html) 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.

#### Basic statistics for this data set

Number of instances in data = 199523

Duplicate or conflicting instances : 46716

Number of instances in test = 99762

Duplicate or conflicting instances : 20936

Number of attributes = 40 (continuous : 7 nominal : 33)

## Data Format

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.