**84.US Census Data**

1. 数据库网址

https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)

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 |  |  | 2458285 |

|  |  |
| --- | --- |
| **Abstract**: The USCensus1990raw data set contains a one percent sample of the Public Use Microdata Samples (PUMS) person records drawn from the full 1990 census sample. |  |

**Source:**

The USCensus1990raw data set was obtained from the (U.S. Department of Commerce) Census Bureau website using the Data Extraction System. This system can be found at <http://dataferrett.census.gov/>. 

**Data Set Information:**

The data was collected as part of the 1990 census.   
  
There are 68 categorical attributes. This data set was derived from the USCensus1990raw data set. The attributes are listed in the file USCensus1990.attributes.txt (repeated below) and the coding for the values is described below. Many of the less useful attributes in the original data set have been dropped, the few continuous variables have been discretized and the few discrete variables that have a large number of possible values have been collapsed to have fewer possible values.   
  
More specifically the USCensus1990 data set was obtained from the USCensus1990raw data set by the following sequence of operations;   
  
- Randomization: The order of the cases in the original USCensus1990raw data set were randomly permuted.   
- Selection of attributes: The 68 attributes included in the data set are given below. In the USCensus1990 data set we have added a single letter prefix to the original name. We add the letter 'i' to indicate that the original attribute values are used and 'd' to indicate that original attribute values for each case have been mapped to new values (the precise mapping is described below).   
  
Hierarchies of values are provided in the file USCensus1990raw.coding.htm and the mapping functions used to transform the USCensus1990raw to the USCensus1990 data sets are giving in the file USCensus1990.mapping.sql.   
  
The data is contained in a file called USCensus1990.data.txt. The first row contains the list of attributes. The first attribute is a caseid and should be ignored during analysis. The data is comma delimited with one case per row.

**Attribute Information:**

--------------------------------------------------------------   
Old Variable New Variable   
--------------------------------------------------------------   
Age dAge   
Ancstry1 dAncstry1   
Ancstry2 dAncstry2   
Avail iAvail   
Citizen iCitizen   
Class iClass   
Depart dDepart   
Disabl1 iDisabl1   
Disabl2 iDisabl2   
English iEnglish   
Feb55 iFeb55   
Fertil iFertil   
Hispanic dHispanic   
Hour89 dHour89   
Hours dHours   
Immigr iImmigr   
Income1 dIncome1   
Income2 dIncome2   
Income3 dIncome3   
Income4 dIncome4   
Income5 dIncome5   
Income6 dIncome6   
Income7 dIncome7   
Income8 dIncome8   
Industry dIndustry   
Korean iKorean   
Lang1 iLang1   
Looking iLooking   
Marital iMarital   
May75880 iMay75880   
Means iMeans   
Military iMilitary   
Mobility iMobility   
Mobillim iMobillim   
Occup dOccup   
Othrserv iOthrserv   
Perscare iPerscare   
POB dPOB   
Poverty dPoverty   
Pwgt1 dPwgt1   
Ragechld iRagechld   
Rearning dRearning   
Relat1 iRelat1   
Relat2 iRelat2   
Remplpar iRemplpar   
Riders iRiders   
Rlabor iRlabor   
Rownchld iRownchld   
Rpincome dRpincome   
RPOB iRPOB   
Rrelchld iRrelchld   
Rspouse iRspouse   
Rvetserv iRvetserv   
School iSchool   
Sept80 iSept80   
Sex iSex   
Subfam1 iSubfam1   
Subfam2 iSubfam2   
Tmpabsnt iTmpabsnt   
Travtime dTravtime   
Vietnam iVietnam   
Week89 dWeek89   
Work89 iWork89   
Worklwk iWorklwk   
WWII iWWII   
Yearsch iYearsch   
Yearwrk iYearwrk   
Yrsserv dYrsserv   
  
Mapping: In this step we map all of the old values for variables with prefix 'd' to new values. The mappings for the variables dAncstry1, dAncstry2, dHispanic, dIndustry, dOccup, dPOB were designed to correspond to a natural coarsening of the original values based on the information in the file coding.htm. The remaining variables are continuous valued variables and the mapping for these variables was chosen to make variables that were fairly uniformly distributed across the states (quantiles). The precise mappings are specified in the file USCensus1990.mapping.sql. This file contains all of T-SQL procedures used to map the variables. These procedures can be used directly in SQLServer to map the original values or translated to some other language.   
--------------------------------------------------------------   
Variable Procedure   
--------------------------------------------------------------   
dAge discAge   
dAncstry1 discAncstry1   
dAncstry2 discAncstry2   
dHispanic discHispanic   
dHour89 discHour89   
dHours discHours   
dIncome1 discIncome1   
dIncome2 discIncome2to8   
dIncome3 discIncome2to8   
dIncome4 discIncome2to8   
dIncome5 discIncome2to8   
dIncome6 discIncome2to8   
dIncome7 discIncome2to8   
dIncome8 discIncome2to8   
dIndustry discIndustry   
dOccup discOccup   
dPOB discPOB   
dPoverty discPoverty   
dPwgt1 discPwgt1   
dRearning discRearning   
dRpincome discRpincome   
dTravtime discTravtime   
dWeek89 discWeek89   
dYrsserv discYrsserv

**Abstract**

The USCensus1990 data set is a discretized version of the [USCensus1990raw](https://archive.ics.uci.edu/ml/machine-learning-databases/census1990-mld/USCensus1990raw.html) data set. Many of the less useful attributes in the original data set have been dropped, the few continuous variables have been discretized and the few discrete variables that have a large number of possible values have been collapsed to have fewer possible values.