Pol Casacuberta Gil
Eric Hurtado
He Chen
Tommaso Patriti
Alexandru-Ilie Popa

Data source: [UCI Machine Learning Repository: Census-Income (KDD) Data Set]

Paragraph explaining how data was obtained:

In order to get the data just use the URL given above and click on Download: Data Folder, you'll be redirected to a different page where you have to click on census.tar.gz. There you'll find three files, the data for the dataset, the explanation of the variables or "features" and the test dataset which we will need to test our machine learning models.
This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau.

Paragraph explaining what data is about:
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.

Nr of records: 299285
nr of variables: 42
nr of numerical variables: 7
nr of categorical variables: 33
Missing values: 415717

Basic structure of data matrix
n: 199523
K: 42
knumeric: 7
kbinary: 2
kquali: 32

Missing per variable:

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
0	0	0	0	0	0	0	0	0	0	0	0	0
V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26
0	0	0	0	0	0	0	0	708	0	0	0	99696
V27	V28	V29	V30	V31	V32	V33	V34	V35	V36	V37	V38	V39
99696	99696	0	99696	0	0	6713	6119	3393	0	0	0	0
V40	V41	V42										
0	0	0										

%missing per variable:

V1	V2	V3	V4	V5	V6	V7	V8
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
V9	V10	V11	V12	V13	V14	V15	V16
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
V17	V18	V19	V20	V21	V22	V23	V24
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.3548463	0.0000000	0.0000000
V25	V26	V27	V28	V29	V30	V31	V32
0.0000000	49.9671717	49.9671717	49.9671717	0.0000000	49.9671717	0.0000000	0.0000000
V33	V34	V35	V36	V37	V38	V39	V40
3.3645244	3.0668144	1.7005558	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
V41	V42						
0.0000000	0.0000000						

Total %missing: 4.96