Entrega D1: US Census Income

Pol Casacuberta Gil

Eric Hurtado

He Chen

Tommaso Patriti

Alexandru-Ilie Popa

Data source: [UCI Machine Learning Repository: Census-Income (KDD) Data Set](https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29)

**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

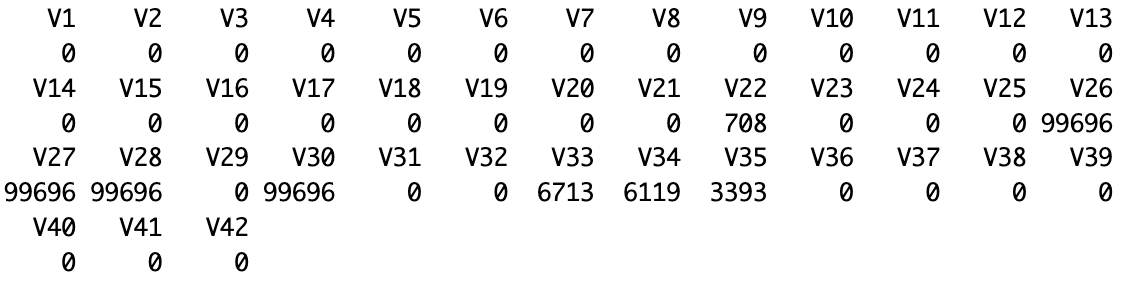
K: 42

knumeric: 7

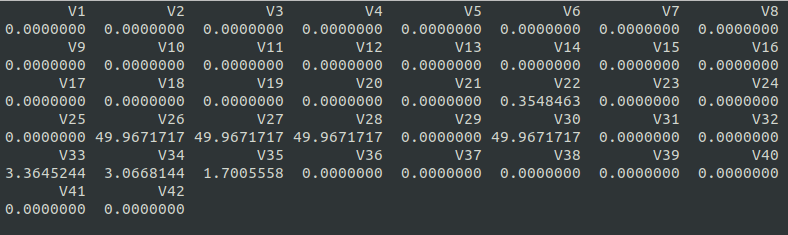
kbinary: 2

kquali: 32

Missing per variable:



%missing per variable:



Total %missing: 4.96