## Case study 1.5 (ZED – TPD 2008)

## Analiza danych o zamożności mieszkańców USA

Ogólnym celem jest ustalenie, które cechy wpływają na przewidywanie stanu zamożności. W ogólności jesteśmy zainteresowani klasyfikacją, czy dochód osoby jest większy, czy mniejszy niż 50 000$.

**Opis danych:**

Dane zebrano w USA w ramach badań demograficznych:

Zbiór danych obejmuje około 25 000 rekordów opisanych klikanastoma zmiennymi. Ostatnia kolumna ma charakter atrybutu decyzyjnego definiującego klasyfikację klientów – zwróć uwagę na niezrównoważenie liczności klas.

Krótka charakterystyka danych podana jest poniżej:

Nazwy zmiennych

age,workclass,demogweight,education,education-num,marital-status,occupation,relationship,race,sex,capital-gain,capital-loss,hours-per-week,native-country,income

Opis ich dziedziny

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Uwagi do analizy:

Warto zwrócić uwagę na

(a) ocena jakości danych i ew. oczyszczanie pliku z niepoprawnych elementów,

(b) redukcja zbioru zmiennych do podzbioru najważniejszych zapewniających satysfakcjonującą ocenę klasyfikacji osób,

(c) poszukiwanie zależności między wartościami wybranych zmiennych a globalną oceną stanu dochodów,

(d) konstrukcja tzw. klasyfikatora, czyli wskazań dla podejmowania końcowej decyzji – porównania wielu sposobów konstrukcji takich klasyfikatorów