Eksploracja danych

Ćwiczenia nr 6

W pliku *Adult\_ch3\_training.csv*dostępne są następujące zmienne:

* *age* – wiek klienta,
* *workclass* – grupa pracownicza (*Federal\_gov* – administracja federalna, *Local-gov* – administracja lokalna, *Never-worked* – nigdy niepracujący*, Private* – sektor prywatny, *Self-emp-inc* – samo zatrudniony (działalność gospodarcza), *Self-emp-not-inc* – samo zatrudniony (bez działalności), *State-gov* – administracja stanowa, *Without-pay* – bez dochodów),
* *education* – lata edukacji,
* *maritalstatus* – stan cywilny (*Divorced* – rozwiedziony, *Married-AF-spouse* – w małżeństwie z osobą z sił zbrojnych, *Married-civ-spouse* – w małżeństwie z cywilem, *Married-spouse-absent* – w małżeństwie z osobą nieobecną, Never married – nigdy niezamężna/nieżonaty, Separeted – w separacji, Widowed – wdowiec/wdowa,
* *occupation* – zawód,
* *sex* – płeć (*Female* – kobieta, *Male* – mężczyzna),
* *capitalgain* – zysk kapitału,
* *capitalloss* – strata kapitału,
* *income* – dochód (*>50K* – ponad 50 tys. dolarów rocznie, *<=50K* – nie więcej niż 50 tys. dolarów rocznie).

Naszym celem jest budowa modelu, który na podstawie pozostałych zmiennych będzie dokonywał klasyfikacji klientów ze względu na dochód. Przeprowadź analizy w programie R.

* 1. Wczytaj plik *Adult\_ch3\_training*.
  2. Dokonaj podziału na zbiory uczący (70%) i testowy (30%).
  3. Zbuduj model metodą drzew C4.5/C5.0, omów go i oceń jego jakość.
  4. Zbuduj las losowy złożony ze 100 drzew CART (pamiętaj o ustawieniu wartości początkowej generatora liczb losowych). Omów go i oceń jego jakość.