

Získávání znalostí z databází
Databáze ze sčítání lidu

1 Úvod

Cílem této práce je získání informací ze zvolené datové sady. V případě tohoto projektu je zvolenou sadou dat databáze ze sčítání lidu.

2 Popis dat

Datová sada se skládá ze 3 souborů - 2 CSV soubory a 1 soubor popisující data (význam jednotlivých sloupců, k čemu data byla použita a jak byla zpracována). CSV soubory jsou sice 2, ale menší soubor obsahuje podmnožinu dat většího a proto se ve výsledku pracuje jen s jedním CSV souborem.

2.1 Popis jednotlivých sloupců

Soubor dat se skládá z celkem z 32561 záznamů a každý záznam z 15 sloupců:

- Age (věk) – celé číslo udávající věk osoby v letech
- Workclass (pracovní třída) – způsob, jakým daná osoba pracuje. Může obsahovat následující hodnoty: *Private*, *Self-emp-not-inc*, *Self-emp-inc*, *Federal-gov*, *Local-gov*, *State-gov*, *Without-pay*, *Never-worked*.
- Final weight (finální hodnota) – celé číslo, pomocí kterého lze určit jak moc podobné si jednotlivé záznamy jsou podobné (čím menší je rozdíl těchto hodnot, tím více si jsou záznamy podobné).
- Education (Vzdělání) – vzdělání, kterého osoba dosáhla. Může obsahovat následující hodnoty: *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 number – Číslo vzdělání??? Možná počet let, kolik studovali??? Číslo ukazující kvalitu vzdělání???
- Marital status (Rodinný stav) – Rodinný stav osoby. Může obsahovat následující hodnoty: *Married-civ-spouse*, *Divorced*, *Never-married*, *Separated*, *Widowed*, *Married-spouse-absent*, *Married-AF-spouse*.
- Occupation (Zaměstnání) – Zaměstnání osoby *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 Vztah/role v rodině – Může obsahovat následující hodnoty:
- Race (Rasa) – Rasová příslušnost dané osoby. Může obsahovat následující hodnoty: *Wife*, *Own-child*, *Husband*, *Not-in-family*, *Other-relative*, *Unmarried*

- Sex (Pohlaví) – údaj, zda se jedná o muže či ženu.
- Capital gain (Kapitálový příjem) – celé číslo udávající
- Capital loss (Kapitálová ztráta) – celé číslo udávající
- Hours per week (Pracovní úvazek) – celé číslo udávající počet hodin, které osoba odpracuje během jednoho týdne
- Native country (Rodná země) – Země, ve které se osoba narodila
- Sallary (Mzda) – Údaj, zda osoba vydělává více jak 50 000 (asi dolarů) ročně.