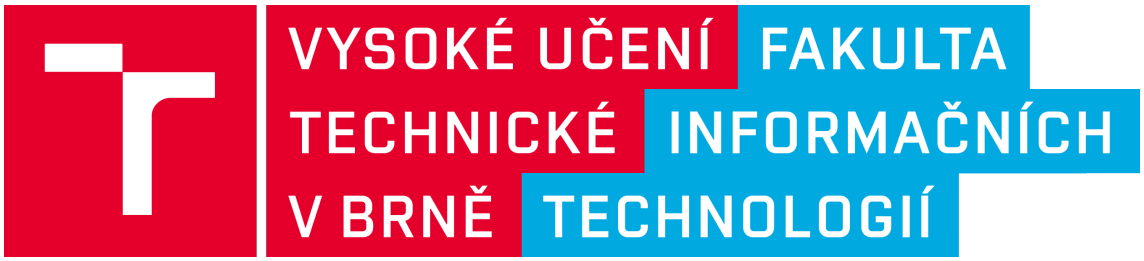

Získávání znalostí z databází
Databáze ze sčítání lidu

20. října 2019

Jiří Matějka, Lucie Pelantová

1 Úvod

Cílem tohoto projektu je získání zajímavých souvislostí z vybrané datové sady. Pro tento účel byla použita data ze sčítání lidu.

Dolováním z této datové sady bude v rámci tohoto projektu zkoumán vliv zázemí občana na výběr jeho povolání. Vytyčené cíle tohoto projektu jsou blíže popsané v kapitole 3.

2 Popis dat

Data ze sčítání lidu byla získána z databáze Census Bureau[1]. Data extrahoval Berry Becker z databáze z roku 1994 a následně z nich pomocí sady podmínek vybral co nejvíce vypovýdající záznamy, mezi tyto podmínky patřilo například omezení na věk občana. Poprvé byla tyto data citována Ronem Kohavi v článku *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid*.

Datová sada se skládá ze 2 částí - soubor **adult.names** obsahuje popis jednotlivých údajů a soubor **adult.data** obsahuje samotné údaje.

3 Cíle projektu

Tento projekt si klade za cíle prozkoumání zajímavých souvislostí mezi zázemím občana a jeho volbou povolání. Při určování zázemí občana budou brány v potaz tyto údaje obsažené v datové sadě:

- Věk
- Dosažené vzdělání
- Rodinný stav
- Role v rodině
- Rasová příslušnost
- Pohlaví
- Rodná země

[1] http://www.census.gov/ftp/pub/DES/www/welcome.html

Vliv na zvolené zaměstnání bude zkoumán ve dvou různých směrech:

- pracovní třída - občan pracuje pod zaměstnavatelem, ve státním sektoru, jako osoba samostatně výdělečně činná, ...
- povolání - občan pracuje například jako prodejce, poskytuje služby, v zemědělství, ...