

Získávání znalostí z databází  
Databáze ze sčítání lidu - zadání

## 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 popsány v kapitole 3.

## 2 Popis dat

Data ze sčítání lidu byla získána z databáze Census Bureau<sup>1</sup>. Data extrahoval Berry Becker z databáze z roku 1994 a následně z nich pomocí sady podmínek (například omezení na věk občana) vybral vypovídající záznamy. Poprvé tyto data citoval Ron 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 samotná data.

## 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

---

<sup>1</sup><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, ...