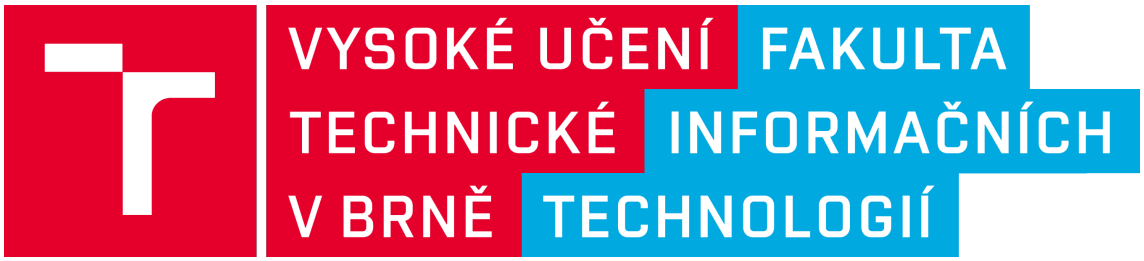

Získávání znalostí z databází
Databáze ze sčítání lidu

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. Záznam o občanu se skládá z těchto údajů:

- Věk
- Pracovní třída
- Finální hodnota
- Dosažené vzdělání
- Počet let studia
- Rodinný stav
- Zaměstnání
- Role v rodině
- Rasová příslušnost
- Pohlaví
- Kapitálový zisk

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

- Kapitálová ztráta
- Výše pracovního úvazku
- Rodná země
- Mzda

3 Cíle projektu

- Pokud chci vydělávat ¿50k, s jakým vzděláním a zaměstnáním mám největší šanci toho dosáhnout? Jak tuto skutečnost ovlivní rodina, rasa a pohlaví?
- Mám vzdělání X, jemi Y let a pracuji Z hodin týdně. Jakou mám šanci, že budu vydělávát ¿50k ročně? Jak tuto pravděpodobnost ovlivní rodina, rasa a pohlaví?
- Nejake dalsi napady?