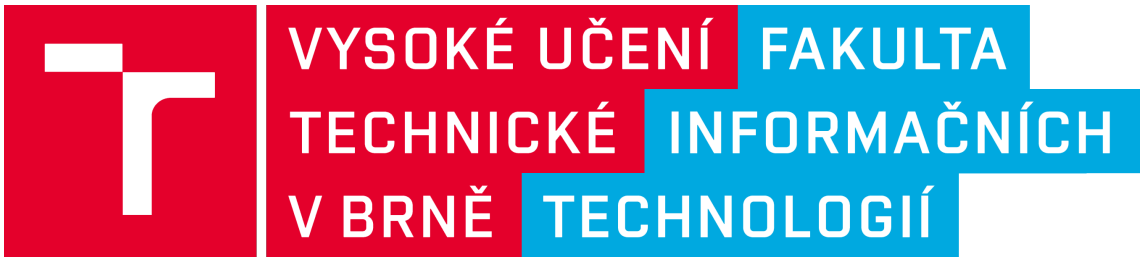

Získávání znalostí z databází
Databáze ze sčítání lidu - zadání

28. listopadu 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 (například omezení na věk občana) vybral vypovýdající 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

[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, ...

4 Předzpracování dat

V části předzpracování byl největší důraz kladen na čištění a redukci dat. Tímto procesem lze dosáhnout kvalitnější analýzi dat v následujících krocích.

4.1 Čištění dat

Jak již bylo zmíněno v kapitole 3, tento projekt je zaměřen na volbu povolání jednotlivých osob. Proto se v atributech týkajících se povolání (pracovní třída a povolání) doplňujeme chybějící hodnoty jejich predikcí. Predikce hodnot je prováděna za pomocí modelu, který byl vytrénovaný na použité datové sadě s použitím algoritmu k-NN.

4.2 Redukce dat

Redukce je dosažena pomocí snížní počtu dimenzí v datové sadě. Dimenze je snížena díky jednoduchému odstranění nerelevantních atributů, jako je například počet let studia či příjmy/výdaje z investic.

5 Získávání znalostí

V této fázi projektu je množina dat rozdělena na trénovací a testovací sadu a to v poměru 7:3. Na základě trénovacích dat je vytvořen rozhodovací strom, který na testovacích datech se snaží předpovídat pracovní třídu a povolání.