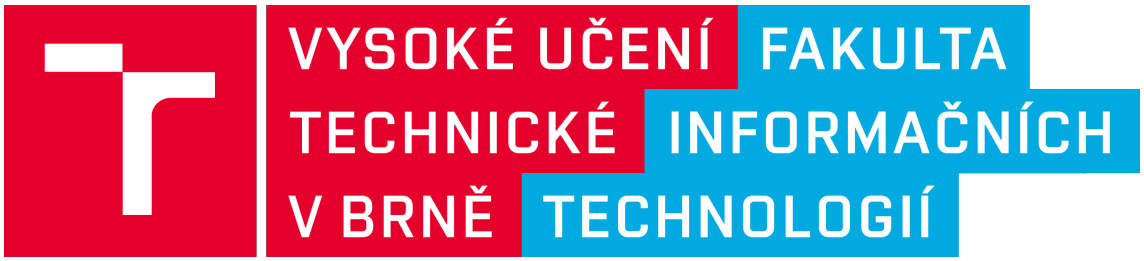


Získávání znalostí z databází
Databáze ze sčítání lidu - zadání

16. prosince 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 byl 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 Návrh projektu

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

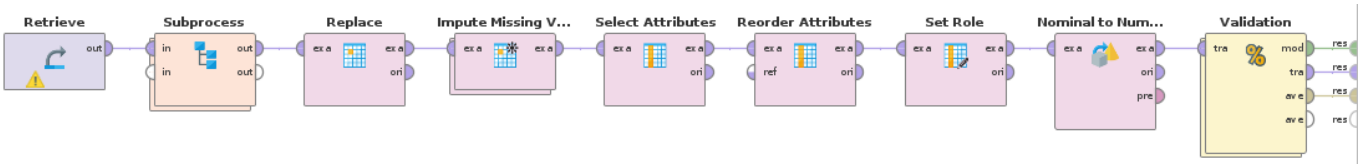
- Věk
- Dosažené vzdělání
- Rodinný stav
- Role v rodině
- Rasová příslušnost
- Pohlaví
- Rodná země

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

Vliv na zvolené zaměstnání byl zkoumán ve dvou různých směrech:

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

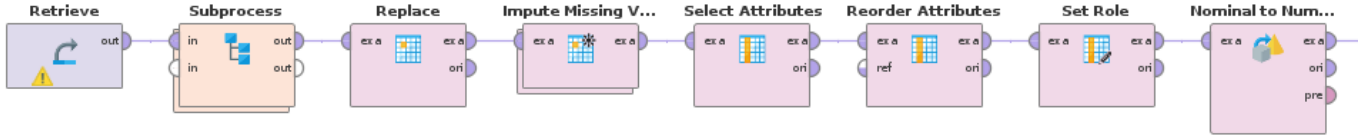
Dolování dat bylo provedeno v programu *rapidminer*. Schéma procesu, který předzpracoval data a následně provedl provedl predixi lze viděť na obrázku 3



Obrázek 1: Výsledné schéma procesu v programu *rapidminer*

4 Předzpracování dat

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



Obrázek 2: Schéma předzpracování v programu *rapidminer*

4.1 Čištění dat

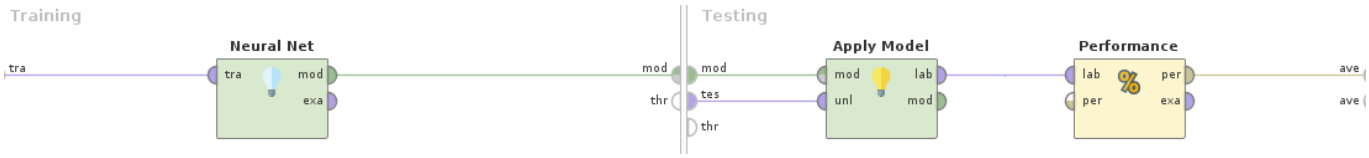
Jak již bylo zmíněno v kapitole 3, tento projekt je zaměřen na volbu povolání jednotlivých osob. Proto se v atributech týkajících se povolání (pracovní třída a zaměření) doplňujeme chybějící hodnoty jejich predikcí. Predikce hodnot je prováděna za pomocí modelu, který byl vytrénovaný na použité datové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

4.2 Redukce dat

Redukce je dosažena pomocí snížení 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 následující části projektu je množina dat rozdělena na trénovací a testovací sadu a to v poměru 7:3, na kterých je model nejprve natrénován a následně je změřena jeho přesnost. V rámci projektu bylo vyzkoušeno několik klasifikačních algoritmů, z níž některé pracují pouze s numerickými daty a proto bylo nutné pro samotnou klasifikaci použít operátor *Nominal to Numerical*, kteý převede původní řetězcové hodnoty na číselné.



Obrázek 3: Schéma aplikace modelu v programu *rapidminer*

5.1 Klasifikační metody

První použitou metodou byl *Naivní Bayesovský klasifikátor*. Jedná se o algoritmus, který dosahuje dobrých výsledků i při trénování na malé množině dat. Na použité datové sadě dosáhl přesnosti 72,93 %. Výsledek dokazuje existenci závislosti atributu **pracovní třídy** na atributech **věk, dosažené vzdělání, rodinný stav, role v rodině, rasová příslušnost, pohlaví a rodné země**.

accuracy: 72.93%

	true State-gov	true Self-emp-not-inc	true Private	true Federal-gov	true Local-gov	true Self-emp-inc	true Without-pay	true Never-worked	class precision
pred. State-gov	2	0	3	1	1	1	0	0	25.00%
pred. Self-emp-not-inc	10	82	248	10	30	45	0	0	19.29%
pred. Private	365	678	7013	273	585	277	4	2	76.25%
pred. Federal-gov	0	0	0	0	0	0	0	0	0.00%
pred. Local-gov	1	4	8	2	15	3	0	0	45.45%
pred. Self-emp-inc	16	12	52	6	6	11	0	0	10.68%
pred. Without-pay	0	0	1	0	0	0	0	0	0.00%
pred. Never-worked	0	0	0	0	0	0	0	0	0.00%
class recall	0.51%	10.57%	95.74%	0.00%	2.35%	3.26%	0.00%	0.00%	

Obrázek 4: Výsledek predikce za použití metody *Naivního Bayovského klasifikátoru*

Další použitou metodou byl *Rozhodovací strom*. Jedná o algoritmus vhodný pro hledání skrytých závislostí v datech. Oproti předchozí metodě (*Naivní Bayesovský klasifikátor*) bylo dosaženo malého zlepšení. Výsledná přesnost tohoto modelu je 74,95 %.

accuracy: 74.95%

	true State-gov	true Self-emp-n...	true Private	true Federal-gov	true Local-gov	true Self-emp-inc	true Without-pay	true Never-wor...	class precision
pred. State-gov	0	0	0	0	0	0	0	0	0.00%
pred. Self-emp-...	0	3	7	0	0	1	0	0	27.27%
pred. Private	394	772	7316	291	637	335	4	2	75.03%
pred. Federal-gov	0	0	0	0	0	0	0	0	0.00%
pred. Local-gov	0	1	1	1	0	0	0	0	0.00%
pred. Self-emp-i...	0	0	1	0	0	1	0	0	50.00%
pred. Without-p...	0	0	0	0	0	0	0	0	0.00%
pred. Never-wo...	0	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.39%	99.88%	0.00%	0.00%	0.30%	0.00%	0.00%	

Obrázek 5: Výsledek predikce za použití *Rozhodovacího stromu*

Zajímavých výsledků bylo dosaženo pomocí *Metody podpůrných vektorů*. Jedná se o často používanou metodu pro klasifikaci a regresní analýzu, protože dosahuje dobrých výsledků pro široké spektrum učících problémů. Metoda dosáhla o něco horších výsledků než *Rozhodovací strom*, přesnost metody byla konkrétně 72,69 %.

accuracy: 72.69%

	true State-gov	true Self-emp-not-inc	true Private	true Federal-gov	true Local-gov	true Self-emp-inc	true Without-pay	true Never-worked	class precision
pred. State-gov	0	0	0	0	0	0	0	0	0.00%
pred. Self-emp-not-inc	0	0	6	0	0	0	0	0	0.00%
pred. Private	361	713	6668	280	613	301	4	2	74.57%
pred. Federal-gov	0	0	0	0	0	0	0	0	0.00%
pred. Local-gov	1	0	4	0	0	0	0	0	0.00%
pred. Self-emp-inc	21	47	115	5	13	32	0	0	13.73%
pred. Without-pay	6	2	16	3	2	2	0	0	0.00%
pred. Never-worked	0	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	97.93%	0.00%	0.00%	9.55%	0.00%	0.00%	

Obrázek 6: Výsledek predikce za použití *Metody podpůrných vektorů*

Z kategorie neuronových sítí byly použity metody *Hluboké učení*, *Neuronové sítě* a *Perceprony*. Tyto metody dosáhly prakticky stejných výsledků jako rozhodovací strom, rozdíly mezi nimi se pohybovaly v řádu desetin procent. Jejich přesnost se pohybovala okolo 74,50 %.

accuracy: 74.64%

	true State-gov	true Self-emp-n...	true Private	true Federal-gov	true Local-gov	true Self-emp-inc	true Without-pay	true Never-wor...	class precision
pred. State-gov	4	2	11	1	3	4	0	0	16.00%
pred. Self-emp-...	3	20	34	4	3	10	0	0	27.03%
pred. Private	385	745	7255	283	622	319	4	2	75.46%
pred. Federal-gov	0	0	0	0	0	0	0	0	0.00%
pred. Local-gov	2	8	18	4	9	2	0	0	20.93%
pred. Self-emp-i...	0	1	7	0	0	2	0	0	20.00%
pred. Without-p...	0	0	0	0	0	0	0	0	0.00%
pred. Never-wo...	0	0	0	0	0	0	0	0	0.00%
class recall	1.02%	2.58%	99.04%	0.00%	1.41%	0.59%	0.00%	0.00%	

Obrázek 7: Výsledek predikce za použití metody *Neuronové sítě*

6 Výsledky

Pro atribut **pracovní třídy** bylo dosaženo nejlepších výsledků (přesnost přibližně 75 %) pomocí metod *Rozhodovacího stromu* a *Neuronových sítí*. Tyto metody byly následně použity i pro predixi atributu **zaměření**, kde byla dosažena přesnost 28,27 % pro metodu *Rozhodovacího stromu* a 29,71 % pro metodu *Neuronové sítě*. Výsledné hodnoty potvrzují existenci souvislosti pracovní třídy osob na základě jejich věku, rasy, pohlaví, vzdělání, místa narození a rodiny. Nicméně v těchto atributech nebyla nalezená dostatečná datová závislost pro predikci samotného zaměření povolání.