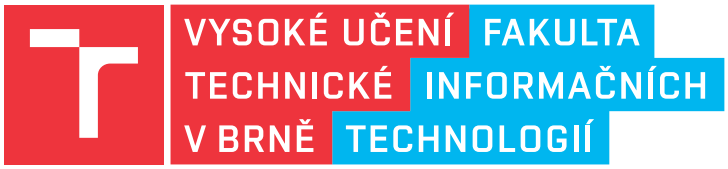


Dokumentace projektu - Řešení
Předmět: **ZZN** 2018/19
Projekt: **Databáze hodnocení hotelů**
Autoři: **Jan Kohút** xkohut08
Michael Halinár xhalin01

1 Úvod

Řešení projektu se zkládá ze dvou hlavních částí, první část se zaměřuje na předpovídání kvality hodnocení. Je v zájmu hotelu reagovat na kvalitní hodnocení za účelem zkvalitnění služeb. Bude použito několik metod pro vytvoření modelu, za účelem získání takového, který bude vykazovat nejlepší výsledky predikce.

Druhou částí projektu je zjišťování asociačních pravidel a vytváření klastrů. Hotel tak může sledovat závislosti mezi spokojeností zákazníků a ostatními faktory pobytu. Při zjištění závsilostí například mezd hodnocením hotelu a ročním období pobytu, je možné se lépe orientovat na konkrétní požadavky a preference zákazníků v daném období. Tyto souvislostí by mohly být potenciálně zajímavé pro vedení hotelu.

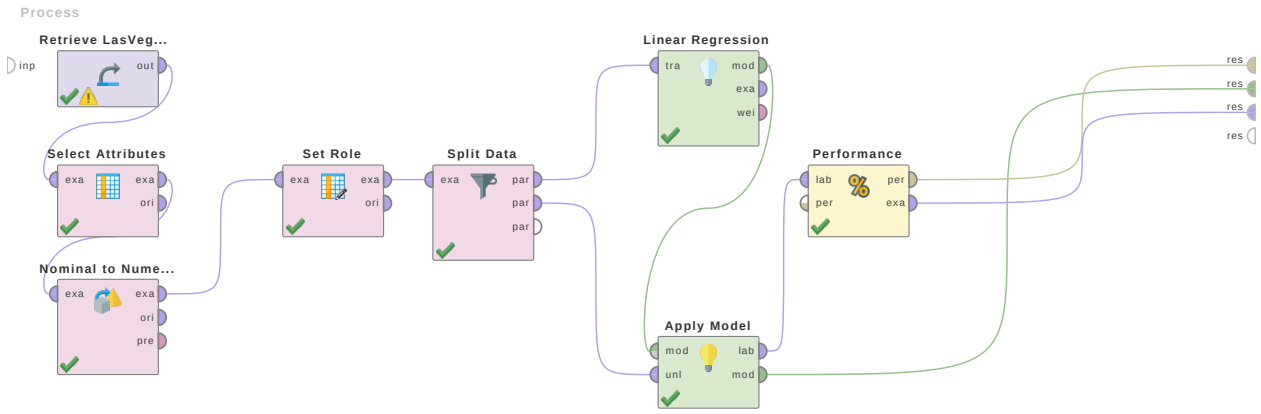
2 Zadání

Popis dat Řádky představují konkrétní hodnocení hotelů. Každý hotel má 24 hodnocení, dvě za měsíc. Ohodnoceno bylo celkem 21 hotelů (tj. 504 hodnocení). Data byla získána ze serveru TripAdvisor, hotely se nachází ve městě Las Vegas. Popis jednotlivých atributů:

user country - země původu hodnotícího uživatele, *nr. reviews* - celkový počet hodnocení uživatele tzn. ne jenom hotely, ale i restaurace, místa atd., *nr. hotel reviews* - celkový počet hodnocení uživatele různých hotelů, *helpful votes* - počet kladných hodnocení (hlasů) na recenzi, *score* - hodnocení hotelu, *period of stay* - doba pobytu uživatele v rozmezích měsíců, *traveler type* - typ pobytu (služební cesta, pár, rodina, sólový pobyt, pobyt s přáteli), *pool, gym, tennis court, spa, casino, free internet* - přítomnost/absence daného zařízení, *hotel name* - název hotelu, *hotel stars* - počet hvězdiček hotelu, *nr. rooms* - počet pokojů hotelu, *user continent* - kontinent ze kterého pochází uživatel, *member years* - počet roků od počátku registrace uživatele na serveru TripAdvisor, *review month* - měsíc ve kterém byla napsána recenze, *review weekday* - den ve kterém byla napsána recenze.

Úloha 1 Klasifikace - Hotel může předpovídat kvalitu hodnocení a na základě toho včas reagovat na dané hodnocení. Při vytvoření nového hodnocení je budoucí počet *helpful votes* neznámý. Pokud má příspěvek v budoucnu velký počet *helpful votes* znamená to, že byl užitečný a kvalitní. Tím pádem je i nejvíce sledován a je žádoucí aby hotel zareagoval na takovýto příspěvek co nejdříve. Cílem je vyzkoušet různé typy klasifikátorů na základě různě velkých podmnožin atributů.

Úloha 2 Clustering a asociační pravidla - V zájmu hotelu, by mělo být zabezpečení vícejazyčného personálu v případě návštěvníků z jiných zemí. Informace o typu a období pobytu je taktéž žádoucí. Hotel se tak může připravit na potencionální zákazníky například kuchyní nebo vhodností pro rodiny



Obrázek 1: Proces lineární regrese v prostředí Rapid Miner.

s dětmi. Cílem je pomocí clusteringu a asociačních pravidel odhalit některé opakující se vzory např. typický původ návštěvníka daného hotelu, závislost důvodu pobytu na období pobytu a zemi původu, vliv období pobytu na hodnocení apod. Hotel se na základě zjištěných vztahů může lépe orientovat na potřeby určitých zákazníků. Pokud bude například zjištěna závislost období pobytu na hodnocení, znamená to nejspíš potenciální problémy v rámci hotelu v obdobích s nižším hodnocením.

3 Řešení

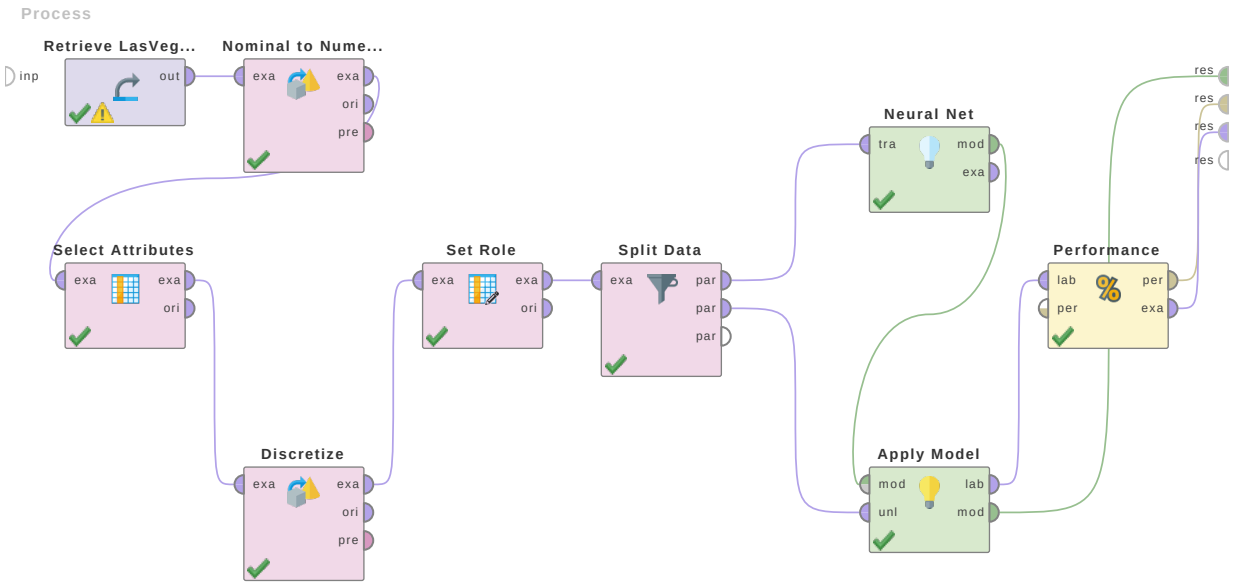
3.1 Úloha 1

Podle korelační matice je patrná závislost *helpful votes* na *nr. reviews* a *nr. hotel reviews*. Tato závislost v podstatě znamená, že zkušenější uživatelé, tzn. ti kteří poskytly více hodnocení, poskytují zpravidla kvalitnější hodnocení. Při snaze určit počet *helpful votes* využíváme závislé atributy a zbytek ignorujeme. Databáze je rozdělena na 403 trénovacích a 100 testovacích hodnocení. Výsledky jsou uváděny pro testovací sadu.

Regrese Jako první přístup jsme vyzkoušeli lineární regresi. Proces v prostředí Rapid Miner je znázorněn na Obrázek 1. RMSD při použití jednotlivých atributů a jejich kombinaci je uvedeno v Tabulka 1. Je patrné, že oba atributy pozitivně přispívají k predikci. Vliv ostatních atributů je prakticky nulový. Většina hodnot atributu *helpful votes* se nachází v rozsahu 0 - 50, zajímavé hodnocení mající přes 100 hlasů jsou zastoupneny v značně menší míře. Predikce je tedy poněkud nepřesná, ale plní požadovaný účel tzn. u zajímavých hodnocení je predikce znatelně vyšší (oproti nezajímavým) byť se značnou odchylkou od reálné hodnoty. Jako druhý přístup jsme vyzkoušeli klasifikaci.

	RMSD
nr.reviews	28.031
nr. hotel reviews	38.360
nr. reviews + nr. hotel reviews	25.342

Tabulka 1: Root mean square deviation při použití různých atributů.



Obrázek 2: Proces klasifikátoru v prostředí Rapid Miner.

Klasifikace Regrese není z uživatelského hlediska příliš pohodlná na použití. Je nutné zkoumat výsledky a na jejich základě vyhodnotit zajímavost/nezajímavost hodnocení. Klasifikátor o dvou třídách představuje jednoduší a přesnější řešení. Odpadá nutnost rozhodovat se na základě odhadnutých hodnot, výsledkem je přímé rozhodnutí o zajímavosti/nezajímavosti. Proces v prostředí Rapid Miner je znázorněn na Obrázek 2. Jako zajímavé hodnocení jsme zvolili hodnocení s více než 65 *helpful votes*. Jako klasifikátor jsme natrénovali neuronovou síť o třech skrytých vrstvách o velikostech 40, 20 a 10. Počet iterací byl 800 a *learning rate* 0.01. Tabulka 2 představuje výsledky pro klasifikaci na základě jednotlivých atributů a jejich kombinace. Opět lze pozorovat lepší výsledky při použití obou atributů. Tabulka 3 prezentuje výsledky pro 3 třídy a to pro počet *helpful votes* v rozsahu $0 - 20$, $21 - 65$ a $66 - infinity$. Je zřejmé, že klasifikátor má problémy ze správným přiřazením do prostřední třídy. Toto je dáno tím, že na základě daných atributů je problematické určit výsledný počet *helpful votes* přesně, ale je možné určit výrazně se lišící hodnocení.

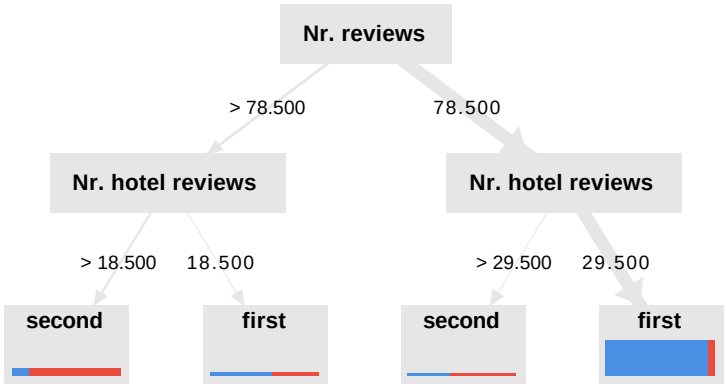
	nezajímavé	zajímavé	celkem
nr.reviews	90.80%	84.62%	90%
nr. hotel reviews	91.57%	70.59%	88%
nr. reviews + nr. hotel reviews	91.95%	92.31%	92%

Tabulka 2: Přesnost klasifikace na základě jednotlivých atributů a jejich kombinace.

Vzhledem k jednoduchosti úlohy je úspěšnost rozhodovacího stromu a naivního bayesovského klasifikátoru podobná. Procesy v prostředí Rapid Miner vypadají stejně jako v případě neuronové sítě, jen je vyměněn model. Příklad rozhodovacího stromu s úspěšností 92% o maximální hloubce 3 je na obrázku 3.

	$0-20$	$21-65$	$66-infinity$	celkem
nr.reviews	80.00%	64.00%	83.33%	76.24%
nr. hotel reviews	73.61%	61.11%	81.82%	72.28%
nr. reviews + nr. hotel reviews	80.30%	64.29%	100%	77.23%

Tabulka 3: Přesnost klasifikace na základě jednotlivých atributů a jejich kombinace (3 třídy).



Obrázek 3: Rozhodovací strom pro úlohu zajímavých hodnocení.

3.2 Úloha 2

Shluková analýza Prvním krokem shlukové analýzy je příprava dat. Pro tento experiment budou použity informace o ročním období pobytu a uděleným hodnocením hotelu. Při tomto výběru dat je teoreticky možné zjistit závyslost mezi ročním obdobím a hodnocením, hotel se tedy může zaměřit na dané roční období a nedostatky které zřejmě přináší, za účelem zlepšení služeb. Obrázek ref fig: clustering ukazuje přípravu dat i samotné shlukování. Po vybrání chtěných atributů následuje převod nominálních hodnot na numerické. Po normalizaci probíhá samotná shluková analýza algoritmem K-Means.

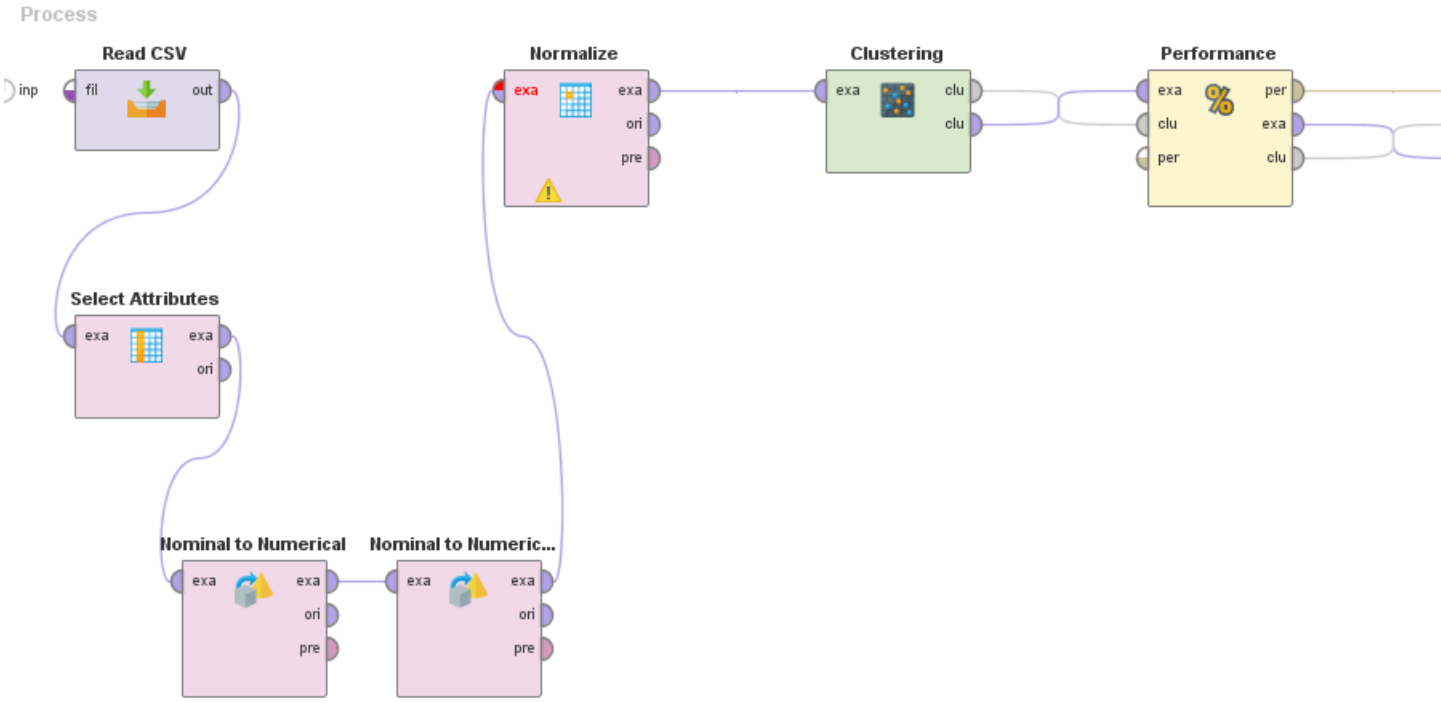
Tento model vytvořil celkem 5 shluků. Obrázek ref fig: clusters ukazuje výsledek analýzy v podobě klusterů s počty 180, 142, 41, 61 a 80 položek, přičemž můžeme vidět že klustery 3 a 4 obsahují hodnocení pouze striktně z jednoho období přitom hotnocení je nízké.

Zajímavý je klastr s číslem 4. Tento obsahuje data pouze s hodnocením z pobytu v období září až listopad. Nejnižší hodnocení hotelu byly zaznamenávány právě v tomto období. Kdy se hodnocení hotelu pohybuje kolem 0.25.

Klastr číslo 3 neobsahuje významné informace, pokud by byl seskupený s klastrem č. 1. nemělo by to příliš velký dopad na celkové průměrné hodnocení v takovém seskupení. Vznikl by tak kluster který by obsahoval hodnocení ze všech období s hodnotou score okolo 1,7 bodů.

Klastr č. 0 a 2 ukazují že vzhledem na výkyvy hodnocení v období září-listopad je stále možné že hotel dostane značně nízké (klastr č.0) respektive poměrně vysoké (klastr č.2) hodnocení.

Asociační pravidla Zvolený dataset hotelů neposkytuje příliš mnoho dat ani větší množství atributů. Proto asociační pravidla budou vytvořeny pouze orientačně pro seznámení se s tímto procesem. Vytvářeny budou jednodimenzionální asociační pravidlá. Medzi zvolené atributy z nichž budou vytvářeny patří položky Casino, Free internet, Gym, Pool, Spa a Tennis court. Jako algoritmus byl použit



Obrázek 4: Model shlukové analýzy v prostředí Rapid Miner.

jediný dostupný v programu RapidMiner a tím je FP-growth.

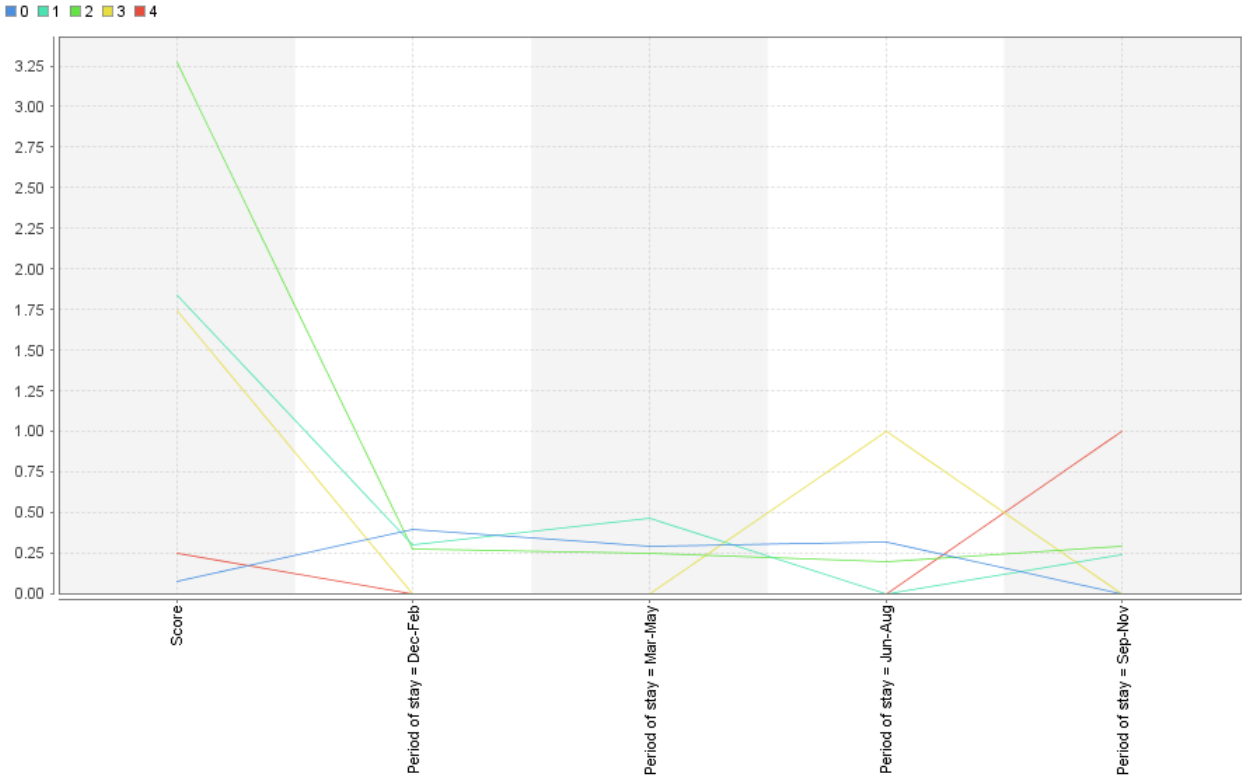
Obrázek ref fig: rules ukazuje vytvořené asociační pravidla. Uvažujme uživatelem zadanou hodnotu podpory a spolehlivosti 50 %. Jediné pravidlo které vyhovuje těmto kritériím je texttt Spa $rightarrow$ texttt Pool, respektive texttt Pool $rightarrow$ texttt Spa s menší hodnotou spolehlivosti. Takový výsledek se dal očekávat když uvážíme že bazén je komponentou lázní.

4 Závěr

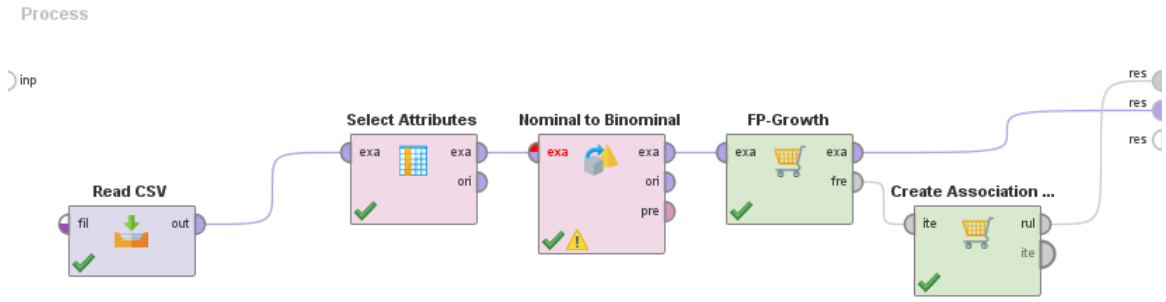
První část projektu se zaměřuje na odhad kvality hodnocení hotelů. Pokud je hodnocení kvalitní tzn. má/bude mít hodně *helpful votes* je pro hotel potenciálně významné a včasná reakce na toto hodnocení je žádoucí. Pro odhad jsme zvolily atributy vyjadřující celkový počet hodnocení a celkový počet hodnocení hotelů, které uživatel zveřejnil. Tyto atributy do značné míry korelují s kvalitou hodnocení. Regrese se ukázala jako vhodná metoda odhadu, ovšem její použití je uživatelsky nepohodlné. Proto jsme diskretizovaly atribut *helpful votes* do více tříd a vyzkoušeli různé klasifikátory a to neuronovou síť, rozhodovací strom a naivní bayes. Úspěšnost těchto klasifikátorů je vzhledem k jednoduchosti úlohy velmi podobná a v nejlepších případech přesahuje 90%. Klasifikátory se tedy jeví jako velmi vhodné pro řešení této úlohy.

Ve druhé části projektu jsme zkusili shlukovou analýzu. Shluky byly vytvářeny na základě atributů období pobytu a hodnocení hotelu. Podařilo se vytvořit jeden shluk který ukazuje že velmi nízké hodnocení hotelů byly udávány v době září až listopad. Díky tomuto zjištění by se hotely měly zaměřit na zvýšení kvality služeb právě v tomto období.

Dalším experimentem bylo vytváření asociačních pravidel. V případě tohoto datasetu vytváření



Obrázek 5: Výsledek shlukové analýzy.



Obrázek 6: Model pro vytváření asociačních pravidel.

takových pravidel nemá až takový smysl jako například při datasetu objednávek e-shopů. Proto byl tento experiment vytvořen orientačně za účelem seznámení se s takovou úlohou. Podařilo se vytvořit jedno rozumné pravidlo které však nemá významnou vypovídající hodnotu.

Premises	Conclusion	Support ↓	Confidence
Pool	Spa	0.762	0.800
Spa	Pool	0.762	1
Tennis court	Pool	0.238	1
Tennis court	Spa	0.190	0.800
Tennis court	Pool, Spa	0.190	0.800
Pool, Tennis court	Spa	0.190	0.800
Spa, Tennis court	Pool	0.190	1
Casino	Pool	0.095	1
Free internet	Pool	0.048	1
Gym	Pool	0.048	1
Free internet	Spa	0.048	1
Free internet	Pool, Spa	0.048	1
Pool, Free internet	Spa	0.048	1
Spa, Free internet	Pool	0.048	1
Tennis court, Casino	Pool	0.048	1

Obrázek 7: Tabulka vytvořených pravidel.