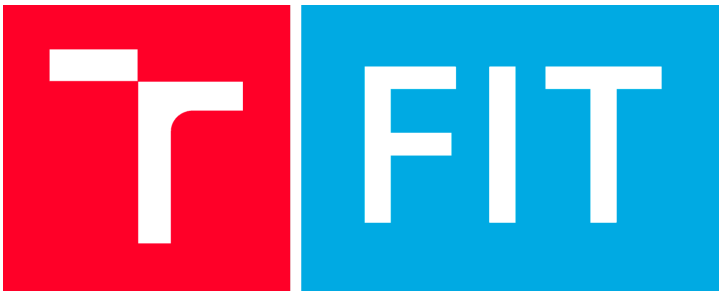
VYSOKÉ UČENIE TECHNICKÉ V BRNE

Fakulta informačných technológií



Získavanie znalostí z databází

2017/2018

Databáza reštaurácií - riešenie

Filip Gulán (xgulan00)

Marek Marušic (xmarus05)

30.11.2017

Obsah

Úvod

Úlohou tohoto projektu bolo vykonať dolovanie dát nad zvolenou dátovou sadou a získať nejaké zaujímavé informácie, ktoré z dát nie sú na prvý pohľad jasné. Ako dátová sada bola zvolené databáza reštaurácií, ktorá je bližšie predstavená a popísané v kapitole *Popis dát*.

V rámci projektu boli riešené 2 samostatné úlohy. Prvá úloha obsiahnutá v rovnomennej kapitole sa týka predikcie spokojnosti zákazníka s reštauráciou na základe informácií o reštaurácii a zákazníkovi. Sú v nej popísané skúšané metódy a postupy, ktoré viedli k potenciálne najlepšiemu výsledku. Metódy sú následne porovnané a sú z nich vyvodené jasné závery. Druhá úloha obsiahnutá v kapitole *Úloha 2* sa týka zhlukovania dát, kedy sa reštaurácia X rozhodla usporiadať večierok a potrebuje zistiť, pre akú skupinu ľudí bude večierok určený. Podľa toho musí reštaurácia vytvoriť čo najvhodnejšie podmienky a ľuďom, ktorý týmto podmienkam vyhovujú je potrebné odoslať pozvánky. Znovu sú v tejto kapitole popísané použité metódy, ich výsledky a následne je popísaná získaná vedomosť.

Popis dát

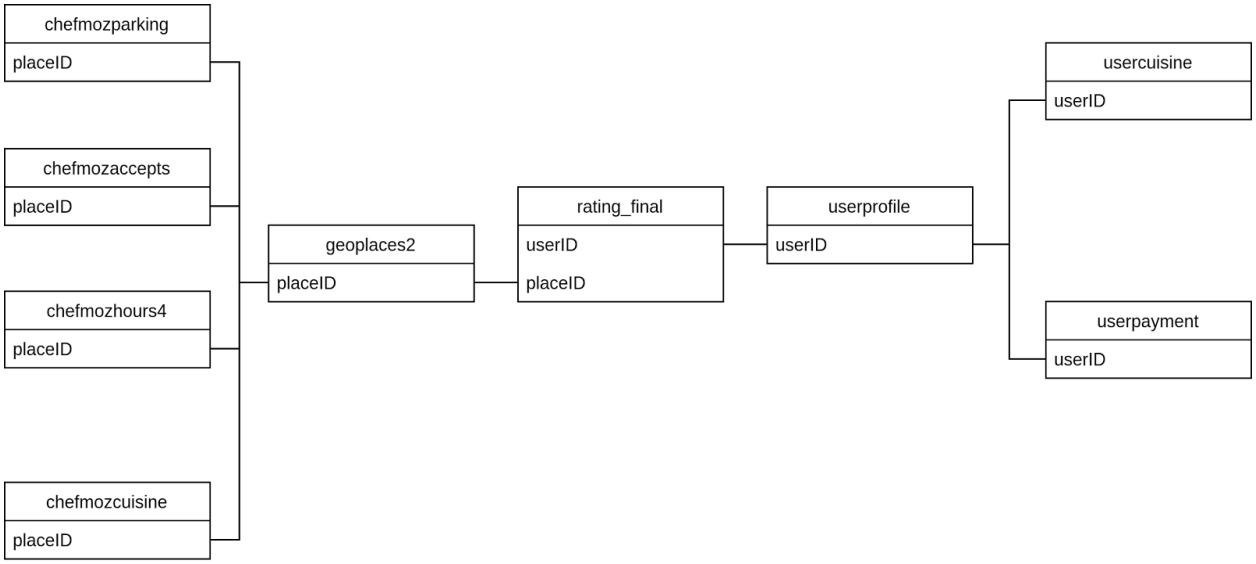
Pre účely tohoto projektu bola zvolená databáza reštaurácií a zákazníkov pre hodnotiaci systém. Autori tejto databázy sú *Rafael Ponce Medellín* a *Juan Gabriel González Serna* z *katedry informatiky Národného Centra Výskumu a Technologického vývoja*[1] (*CENIDET*) v *Mexiku*. V minulosti bola databáza použitá pre vedecký článok *Effects of relevant contextual features in the performance of a restaurant recommender system*[2] a pre zostavenie zoznamu top-n reštaurácií podľa preferencií zákazníkov.

Databáza obsahuje 9 *CSV* súborov, ktoré obsahujú dáta o reštauráciach, zákazníkoch a o hodnoteniach reštaurácií zákazníkmi. Súbory databázy sú nasledujúce:

1. Informácie o reštauráciach
    a. *chefmozaccepts.csv* - podporované typy platieb v reštauráciach (id reštaurácie, typ platby).
    b. *chefmozcuisine.csv* - typy kuchýň v reštauráciach (id reštaurácie, typ kuchyne)
    c. *chefmozhours4.csv* - otváracie hodiny reštaurácií (id reštaurácie, otváracie hodiny, dni v týždni).
    d. *chefmozparking.csv* - typy parkovacích miest (id reštaurácie, typ parkovacieho miesta).
    e. *geoplaces2.csv* - reštaurácie (id reštaurácie, zemepisná šírka, zemepisná dĺžka, meno reštaurácie, adresa, mesto, štát, krajina, fax, PSČ, možnosť pitia alkoholu, fajčiarsky priestor, dress code, prístupnosť, cenová kategória, url webovej stránky, atmosféra, posedenie vonku, ostatné služby).
2. Informácie o zákazníkoch
    a. *usercuisine.csv* - typy obľúbených kuchýň zákazníka (id zákazníka, typ kuchyne).
    b. *userpayment.csv* - platobné možnosti zákazníka (id zákazníka, typ platby).
    c. *userprofile.csv* - zákazníci (id zákazníka, zemepisná šírka, zemepisná dĺžka, vzťah k fajčeniu, vzťah k alkoholu, obľúbené oblečenie, preferovaná atmosféra, typ prepravy, stav, potomstvo, rok narodenia, záujem, osobnosť, vierovyznanie, aktivita, farba, hmotnosť, rozpočet, výška).
3. Hodnotenie reštaurácií
    a. *rating_final.csv* - hodnotenia reštaurácií zákazníkmi (id zákazníka, id reštaurácie, hodnotenie, hodnotenie jedla, hodnotenie služieb).

Vzťahy medzi súbormi (tabuľkami) databázy je možné vidieť v jednoduchej schéme vzťahov na obrázku 1. Pre názornosť boli v tabuľkách zobrazené iba atribúty, ktoré sa podieľajú na vzťahu medzi tabuľkami a všetky ostatné atribúty boli vynechané.

[1] https://www.cenidet.edu.mx/
[2] http://ceur-ws.org/Vol-791/paper8.pdf



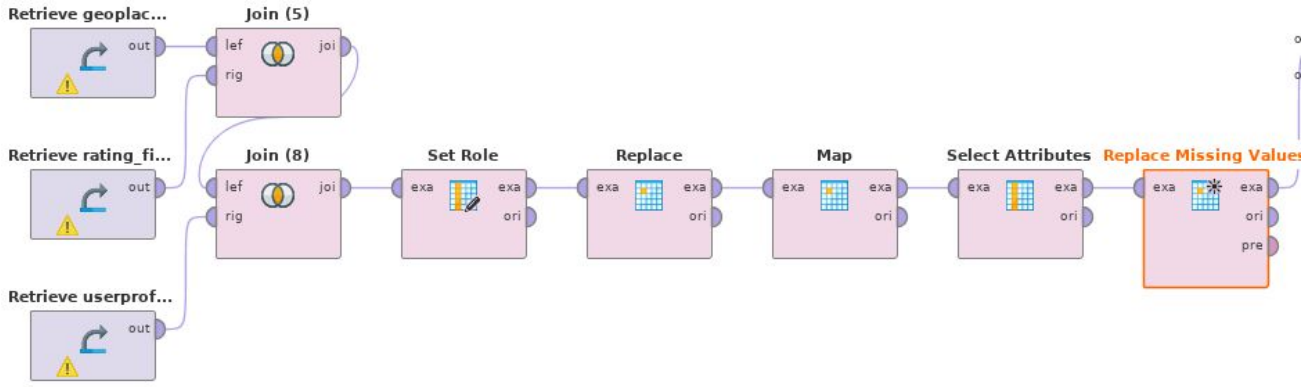
Obrázok 1: Schéma vzťahov medzi tabuľkami databázy.

Úloha 1

Úloha 1 sa zaoberá predpoveďou spokojnosti zákazníka s reštauráciou na základe informácií o reštaurácii a zákazníkovi. Podobnú predpoveď je možné využiť napr. v systéme, ktorý by mohol zákazníkom cielene odporúčať reštaurácie, s ktorými budú s vysokou pravdepodobnosťou veľmi spokojní.

Príprava dát

Dáta boli poskytnuté vo formáte niekoľkých *CSV* súborov. Po preskúmaní obsahu dát a experimentovaní s rôznymi atribútami zo všetkých *CSV* súborov boli pre účel tejto úlohy zvolené 3 súbory s názvami *geoplaces2.csv*, *rating_final.csv* a *userprofile.csv*. Pri importovaní dát boli nastavené správne datové typy. Pred spojením tabuliek bolo nutné premenovať niektoré atribúty, keďže viacero tabuliek obsahovalo atribúty s rovnakým názvom. Potom bolo možné tabuľky spojiť na základe *userID* a *placeID*. Atribút *rating* bol neskôr nastavený ako *label*, pretože sa budeme snažiť predpovedať jeho hodnotu. Po experimentovaní boli zvolené atribúty pri ktorých modely dosahovali najlepšie výsledky. Taktiež bolo nutné vyčistiť dáta od nekonzistencií ako napríklad rôzne mená pre dané mesto, kedže napr. mesto *San Luis Potosi* bolo pomenované 5 rôznymi názvami resp. skratkami. Nakoniec boli doplnené chýbajúce hodnoty priemernou hodnotou (viď. Obrázok 2).

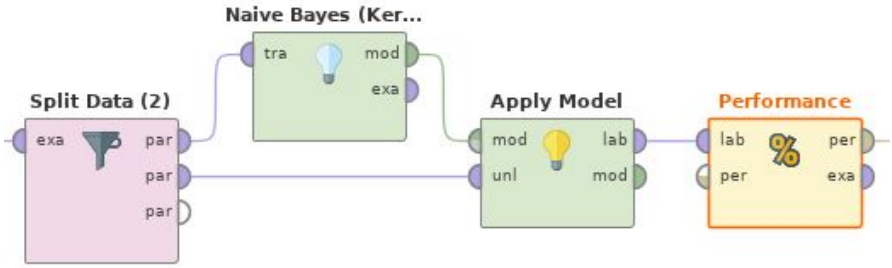


Obrázok 2: Príprava dát.

Dolovacie metódy

V tejto dolovacej úlohe budú použité metódy pre klasifikáciu do 3 tried. Tieto triedy predstavujú hodnotenie (atribút *rating*) reštaurácie daným návštevníkom reštaurácie. Hodnotenie bolo zaznamenané pomocou hodnôt 0, 1 ,2 kde 0 je najväčšia nespokojnosť a 2 je najväčšia spokojnosť. Pred samotným použitím modelov boli dáta rozdelené na trénovaciu sadu (*90%* dát) a testovaciu sadu (*10%* dát). Pri rozdelení bolo použité *stratifikované* rozdelenie vzoriek, čo

zabezpečí že náhodne rozdelí dáta tak aby testovacia a dátova sada obsahovali triedy rozložené približne rovnako ako v celej dátovej sade. Pre validáciu dát bol model aplikovaný pomocou *apply model* a presnosť bola meraná pomocou *performance* operátoru (viď Obrázok 3).



Obrázok 3: Validácia modelov.

Naive Bayes

Po experimentovaní s rôznymi modelmi a atribútami bol zvolený ako prvý model *Naive Bayes*. Najskôr boli parametre zvolené postupným experimentovaním so selekciou rôznych parametrov čo dospelo k výsledku *66.38%* (Obrázok 4). Zvolené boli nasledovné parametre: *alcohol, city, dress_code, dress_preference, drink_level, marital_status, placeID, rating, religion, userID.* Tieto parametre budú používané aj v daľších modeloch ako referencia.

accuracy: 66.38%

	true 2	true 1	true 0	class precision
pred. 2	35	12	6	66.04%
pred. 1	9	27	4	67.50%
pred. 0	5	3	15	65.22%
class recall	71.43%	64.29%	60.00%	

Obrázok 4: Presnosť klasifikácie *Naive Bayes* metódou.

Neskôr bola zvolená optimalizácia selekcie pomocou evolučných algoritmov v *RapidMiner* čo viedlo k zlepšeniu výsledku na *73.28%* (viď. Obrázok 5). Použitá veľkosť populácie bola 30, počet iterácií taktiež 30 a *random seed* bolo nastavené na 1992. Po optimalizácií boli zvolené tieto parametre: *activity, alcohol, city, color, dress_code, hijos, interest, longitude, marital_status, name, personality, price, smoker, transport, userID, weight.*

accuracy: 73.28%

	true 2	true 1	true 0	class precision
pred. 2	40	7	4	78.43%
pred. 1	7	30	6	69.77%
pred. 0	2	5	15	68.18%
class recall	81.63%	71.43%	60.00%	

Obrázok 5: Presnosť klasifikácie *Navieve Bayes* po optimalizácii selekcie atribútov.

Gradient boosted trees

Pri použití rovnakých atribútov ako pri modeli *Naieve Bayes*, ktoré boli získané ručne bola nameraná presnosť *63.79%* (viď. Obrázok 6).

accuracy: 63.79%

	true 2	true 1	true 0	class precision
pred. 2	32	11	6	65.31%
pred. 1	14	28	5	59.57%
pred. 0	3	3	14	70.00%
class recall	65.31%	66.67%	56.00%	

Obrázok 6: Ručné experimentovanie s atribútmi v modeli *Gradient boosted trees*.

Po aplikovaní optimalizácie selekcie evolučnými algoritmami bola dosiahnutá presnosť *72.41%* (viď. Obrázok 7) pri použití nasledujúcich atribútov: *accessibility, alcohol, ambience, area, birth_year, city, color, dress_code, dress_preference, drink_level, franchise, name, personality, userID, zip*.

accuracy: 72.41%

	true 2	true 1	true 0	class precision
pred. 2	41	12	5	70.69%
pred. 1	7	28	5	70.00%
pred. 0	1	2	15	83.33%
class recall	83.67%	66.67%	60.00%	

Obrázok 7: Optimalizácia selekcie pre *Gradient boosted trees* metódu.

K-nearest neighbors

Pri použití referenčných atribútov bola získaná presnosť *56.03%* (viď. Obrázok 8).

accuracy: 56.03%

	true 2	true 1	true 0	class precision
pred. 2	31	15	6	59.62%
pred. 1	14	22	7	51.16%
pred. 0	4	5	12	57.14%
class recall	63.27%	52.38%	48.00%	

Obrázok 8: *K-NN* presnosť s referenčnými atribútmi.

Po evolučnej optimalizácii atribútov bola získaná presnosť *68.97%* (viď. Obrázok 9) s nasledujúcimi atribútmi: *accessibility, address, ambience, birth_year, color, dress_preference, drink_level, franchise, hijos, interest, latitude, marital_status, personality, smoking_area, the_geom_meter, userID*.

accuracy: 68.97%

	true 2	true 1	true 0	class precision
pred. 2	35	14	3	67.31%
pred. 1	11	26	3	65.00%
pred. 0	3	2	19	79.17%
class recall	71.43%	61.90%	76.00%	

Obrázok 9: *K-nearest neighbors* s optimalziovanými atribútmi.

Deep learning

Hlboká neurónová sieť mala presnosť *56,90%* pri použití referenčných atribútov (viď. Obrázok 10).

accuracy: 56.90%

	true 2	true 1	true 0	class precision
pred. 2	45	35	8	51.14%
pred. 1	2	5	1	62.50%
pred. 0	2	2	16	80.00%
class recall	91.84%	11.90%	64.00%	

Obrázok 10: Hlboká neurónová sieť s referenčnými atribútmi.

Po aplikácii evolučnej optimalizácie atribútov bola zvýšená presnosť na 72.41%. Optimalizovaná selekcia atribútov: *accessibility, alcohol, ambience, area, birth_year, city, color, dress_code, dress_preference, drink_level, franchise, name, personality, userID, zip.*

accuracy: 72.41%

	true 2	true 1	true 0	class precision
pred. 2	41	13	5	69.49%
pred. 1	6	26	3	74.29%
pred. 0	2	3	17	77.27%
class recall	83.67%	61.90%	68.00%	

Obrázok 11: Hlboká neurónová sieť s optimalizovanými atribútmi.

Zhodnotenie

Pri modeloch bola taktiež aplikovaná optimalizácia parametrov pomocou evolučných algoritmov, avšak bohužiaľ vyššiu presnosť sa nepodarilo získať. Príčinou tohto javu môže byť *random seed* pri rozdeľovaní dátovej sady (operátor *Split Data*) čo mohlo zamaskovať pozitívny vplyv zmien parametrov. Najlepší výsledok sa podarilo dosiahnuť pomocou modelu *Naive Bayes* a to *73.28%* presnosť pri klasifikácii s použitými atribútmi *accessibility, activity, alcohol, franchise, height, interest, longitude, other_services, placeID, Rambience, rating, smoker, userID, zip.* Metóda *Gradient Boosted trees* nebola o moc horšia s presnosťou *72.41%*. Najčastejšie boli optimalizácoiu zvolené atribúty *color (4x), personality (4x), userID (4x), accessibility (3x), alcohol (3x), birth_year (3x), city (3x), dress_code (3x), dress_preference (3x), drink_level (3x), franchise (3x), name(3x), ambience (2x), area (2x), hijos (2x), interest (2x), latitude (2x), marital_status (2x), zip (2x).*

Úloha 2

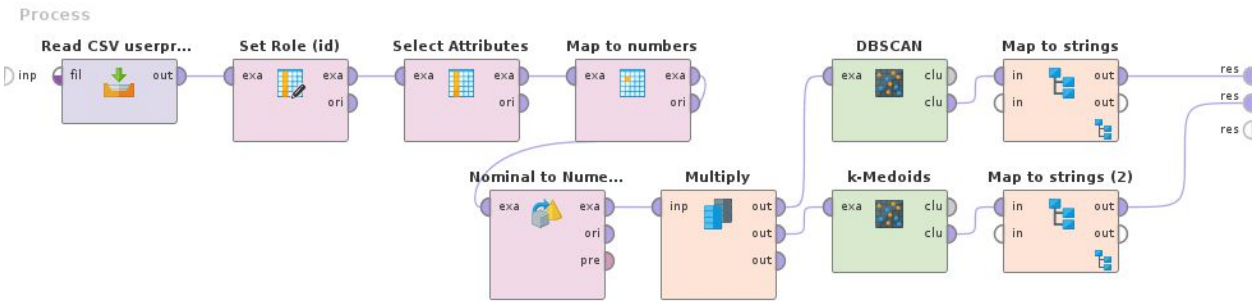
Úloha číslo 2 využíva dolovaciu metódu zhlukovania dát do jednotlivých zhlukov. Zhluky sú špecifické tým, že dáta vnútri zhluku by mali byť navzájom čo najviac k sebe podobné a dáta medzi zhlumi by mali byť zase čo možno najviac odlišné.

Formulácia

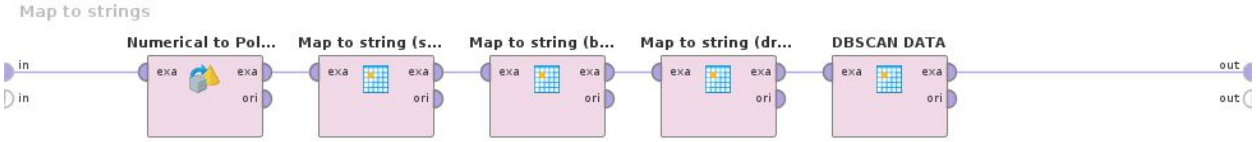
Reštaurácia X sa rozhodla usporiadať večierok a potrebuje zistiť, pre akú skupinu ľudí bude večierok určený. Podľa toho musí reštaurácia vytvoriť čo najvhodnejšie podmienky a ľuďom, ktorý týmto podmienkam vyhovujú je potrebné odoslať pozvánky. Reštaurácia má k dispozícií dáta o minulých návštevníkoch sietí reštaurácií.

Postup

Úloha je riešená pomocou zhlukovania, kedy sa návštevníci priraďujú do zhlukov. Pracovalo sa s tabuľkou *userpayment* a s atribútmi *smoker, drink_level, dress_preference* a *budget*. *CSV* dáta boli najprv načítané z *CSV* súboru do programu *Rapidminer* pomocou bloku *Read CSV*. Dátam bol následne nastavený identifikátor na položku *userID* a boli vybrané stĺpce, s ktorými sa ďalej pracovalo (*smoker, drink_level, dress_preference a budget*). Stĺpce obsahovali nominálny typ dát (reťazce), ktoré boli namapované na nominálny typ dát (čísla) a tieto nominálne číselné hodnoty boli následne prevedené na numerický typ. Mapovanie, napríklad pre atribút *smoker,* bolo vykonané nasledovne.: Hodnota *true* sa namapovala na číslo 2, hodnota *false* na číslo 0 a chýbajúca hodnota na číslo 1. Obdobne sa mapovali aj ostatné atribúty, tj. vždy sa zvolil nejaký rozsah a vždy sa nominálne hodnoty namapovali na tento rozsah. Potenciálne minimum bolo vždy mapované na číslo 0 a potenciálne maximum bolo mapované na najväčšie možné číslo (počet možných hodnôt - 1). Zmyslom tohoto mapovania bolo, aby sa v zhlukoch mohli nachádzať napríklad fajčiari a tí o ktorých nie je známe či sú fajčiari, ale zároveň aby sa tam nenachádzali fajčiari a nefajčiari. Po prevedení mapovania a konverzie na numerický typ, boli dáta zmultiplikované a boli na ne aplikované zhlukovacie metódy. Výstupom zhlukovacích metód boli dáta rozdelené do zhlukov a model, ktorý hovoril o tom, koľko sa v danom zhluku nachádzalo položiek. V rámci úlohy bolo testovaných viacero zhlukovacích metód, z toho sa najefektívnejšie zdali *DBSCAN* a *k-Metoids*. Pre čitateľnosť boli dáta po zhlukovaní prevedené na polynomiálny typ a znovu mapované na originálne hodnoty. Napríklad pre atribút *smoker*, ktorý po všetkých operáciách obsahoval hodnoty 0,1,2 sa hodnota 0 namapovala na *not smoker*, hodnota 1 na *unknown* a hondota 2 na *smoker*. Tento proces mapovania bol vykonaný v *subprocese*. Celý proces získania zhlukov dát je možné vidieť na obrázku 12 a detail subprocesu mapovania dát z numerických hodnôt na reťazcové je možné vidieť na obrázku 13.



Obrázok 12: Proces získavania zhlukov dát

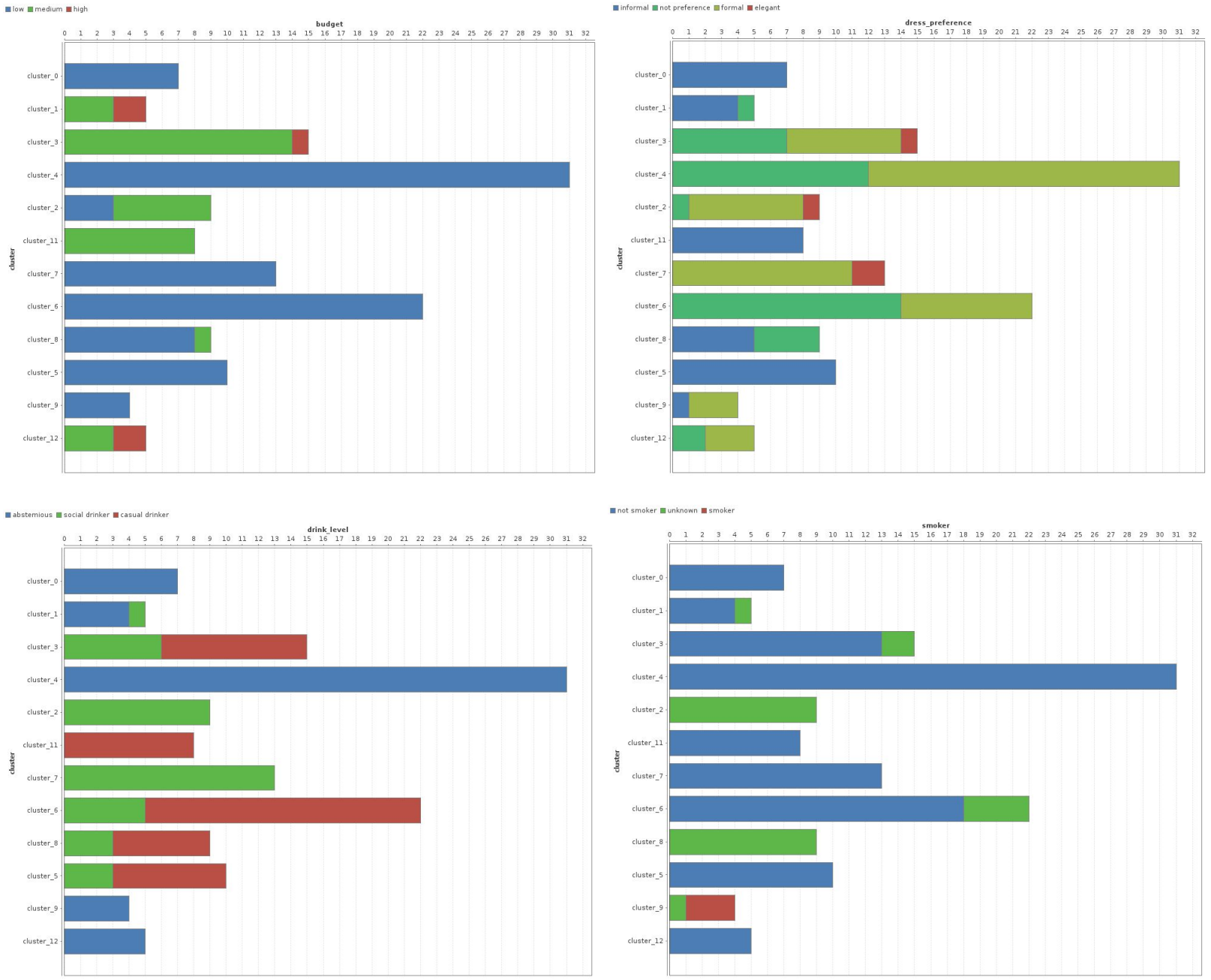


Obrázok 13: Detail subprocesu mapovania dát z numerických hodnôt na reťazcové

Výsledky

Výsledky zhlukovania boli zanesené do stĺpcových grafov, v ktorých je vidieť zastúpenie danej skupiny v danom zhluku a je potom možné na základe tejto informácie robiť určité závery.

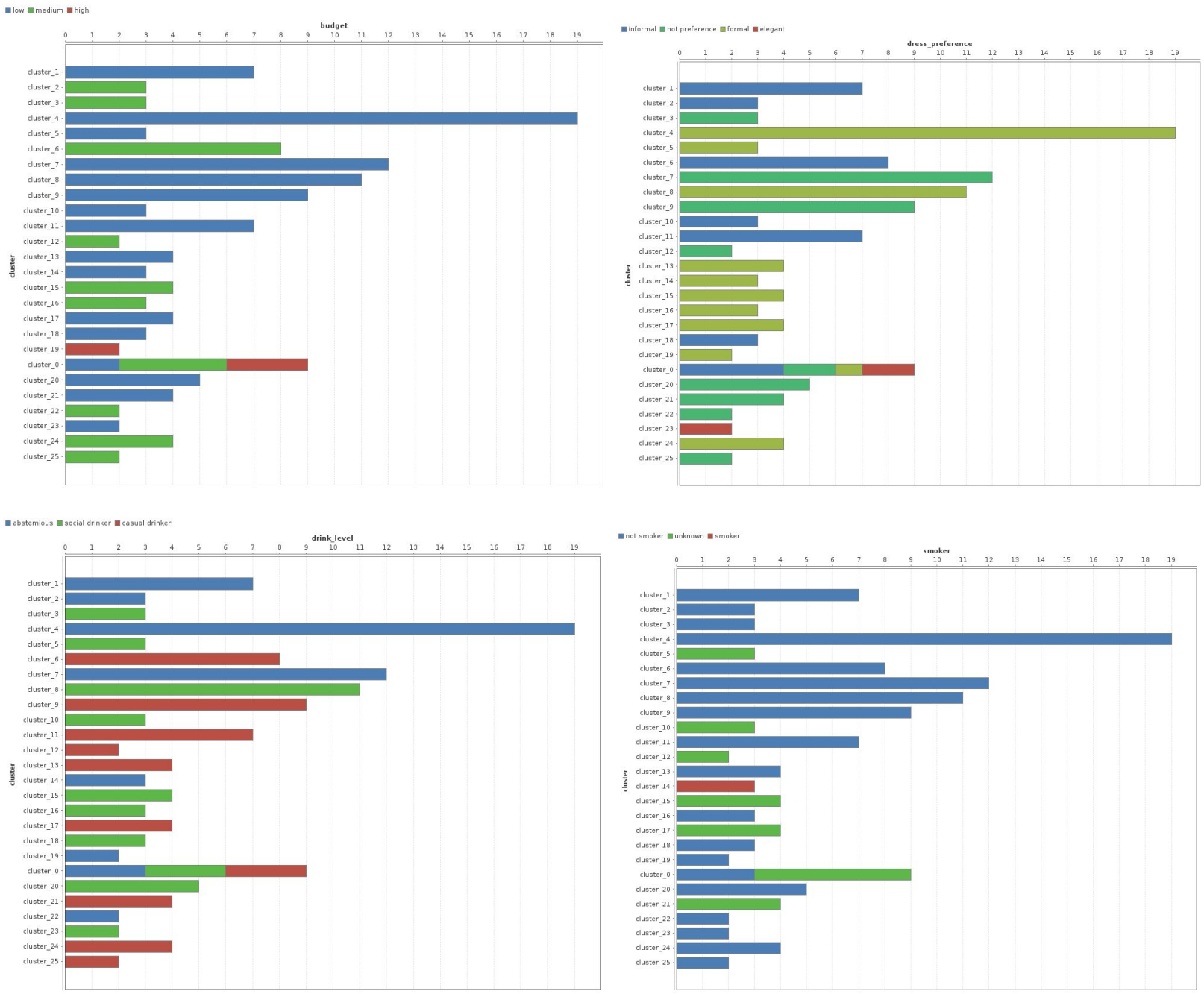
Metódou *k-Medoids (Numerical measure: CosineSimilarity, k: 15, max runs: 15, max optimization steps: 200)* sme získali 12 zhlukov, ktoré obsahujú návštevníkov reštaurácie. Z grafov je vidieť, že najväčší zhluk je zhluk 4, ktorý obsahuje 31 návštevníkov, ktorý sú nefajčiari, preferujú formálne oblečenie alebo nemajú pevne danú preferenciu oblečenia, zároveň ale majú nízky rozpočet a sú abstinenti. Reštaurácia by teda mohla usporiadať večierok v nefajčiarskej zóne, kde by bolo vyžadované formálnejšie oblečenie, podávané jedlo nižšej cenovej kategórie a nebolo by potrebné podávať alkohol. Týmto zistením, by reštaurácia ušetrila náklady na alkohole, keďže nie je potrebný a zároveň na surovinách vyššej cenovej kategórie, keďže by sa podávalo jedlo celkovo nižšej kvality. Najmenší zhluk je zhluk 9, ktorý obsahuje iba 4 návštevníkov, ktorý majú k dispozícii nižší rozpočet, preferujú formálne, alebo neformálne oblečenie, sú fajčiari (alebo nemajú uvedené či sú fajčiari/nefajčiari) a sú abstinenti. Pre takúto malú skupinu ľudí a s takýmito parametrami by nemalo zmysel usporadovať večierok. Grafy z ktorých sú vykonávané závery je možné vidieť nižšie.



Obrázok 14: Výsledky metódy k-Medoids

Metódou *DBSCAN (Numerical measure: EuclideanDistance, epsilon: 1.0, min points: 2)* sme získali 26 zhlukov, ktoré obsahujú návštevníkov reštaurácie. Z grafov je možné vidieť, že najväčší zhluk je zhluk 4, ktorý obsahuje 19 návštevníkov, ktorý sú nefajčiari, preferujú formálne oblečenie, zároveň ale majú nízky rozpočet a sú abstinenti. Závery tejto analýzy sú totožné so závermi analýzy vykonanej pri metóde *k-Medoids*. Najmenej návštevníkov obsahuje hneď niekoľko zhlukov, ktoré obsahujú iba 2 návštevníkov. Tento jav s 2 účastníkmi je kvôli nastavení *DBSCAN*, parametru minimálneho počtu bodov nastaveného na 2. Pri vyššom čísle zhluky

obsahovali nie až tak úplne podobných návštevníkov a bol prítomný stĺpec cca o počtu 20 účastníkov, ktorý predstavoval ľudí, ktorý sa nedali zaradiť do žiadneho ďalšieho zhluku a v rámci daného zhluku sa zdali byť nie veľmi podobný.



Obrázok 15: Výsledky metódy *DBSCAN*

Metóda *DBSCAN* a metóda *k-Medoids* dosahovali veľmi podobné výsledky, aj keď metóda *k-Medoids* sa javí mierne lepšie, keďže našla menej a väčšie zhluky podobných užívateľov.

Záver

V tomto projekte boli riešené dve dolovacie úlohy nad databázou reštaurácií a zákazníkov v prostredí *RapidMiner*. Prvá úloha spočívala v predpovedi spokojnosti zákazníkov s danou reštauráciou. Úloha bola riešená pomocou 4 rôznych klasifikačných modelov, konkrétne *Naive Bayes* s presnosťou *73.28%*, *Gradient boosted trees* s presnosťou *72.41%*, *K-nearest neighbors* s presnosťou *68.97%* a *Deep learning* s presnosťou 72.41%. Výsledky boli získané za pomoci vyčistenia dát a aplikovaním optimalizácie selekcie parametrov.

Ďalšia úloha spočívala v zhlukovaní návštevníkov reštaurácií do skupín za účelom plánovania večierkov pre cielené skupiny užívateľov. Takto môžu reštaurácie prispôsobiť plány večierku pre danú skupinu ľudí. Po porovnaní a experimentovaní s rôznymi modelmi boli vybrané metódy *k-Medoids* a *DBSCAN*. Najlepšie sa darilo metóde *k-Medoids*, ktorej sa podarilo získať najväčší zhluk s 31 návštevníkmi. Získaný návštevníci sú abstinujúci nefajčiari, ktorý preferujú formálne oblečenie a majú nižší rozpočet. Takýmto spôsobom môže reštaurácia pripraviť podmienky, v ktorých bude najväčšia skupina ľudí spokojná a tak isto môže reštaurácia vytvoriť cielenú reklamu na užívateľov z daného zhluku.