Získavanie znalostí z databází 2018/2019

Databáza hodnotenia hotelov
Riešenie

Michal Ormoš (xormos00)
Petra Mikulová (xmikul67)

Obsah

1 Úvod **2**

2 Popis dát **3**

3 Predspracovanie dát **5**

4 Prvotné dolovanie dát **6**

5 Dolovacia úloha **9**
5.1 Formulácia 9
5.2 Postup 9
5.3 Výsledky 10

6 Záver **12**

1 Úvod

Úlohou tohto projektu bolo vykonať dolovanie dát nad zvolenou datovou sadou a získať zaujímavé informácie, ktoré nie su na prvý pohľad hneď viditeľné. Ako datová sada bola zvolená `databáza hodnotenia hotelov` v meste Las Vegas, ktorá bola sprostredkovaná fakultou informačných technológií v Brne. Bližšie bude predstavená v 2. kapitole.

V prvom kroku, v kapitole 3, sme sa venovali predspracovaniu dát, kde sme dáta upravili do formy vhodnej pre ich následne spracovanie. V druhom kroku, kapitole 4, sme sa zoznamovali s dátami a hľadali možné spojitosti medzi nimi, ktoré by mohli byť užitočné, nevenovali sme sa žiadnym zložitejšim dolovacím algoritmom. V poslednom kroku sme riešili zadanie, ktoré sme si určili v prvej časti projektu za pomoci dolovacích algoritmov, a to hlavne Naive Bayes.

Všetky kroky tohto projektu, od spracovania dát, až po ich analýzu, boli spracované za pomoci nástroja `RapidMiner` s použitim jeho študentskej licencie registrovanej na školský email.

2 Popis dát

V tomto projekte sme pracovali s databázou hotelov z mesta Las Vegas, ktorá bola podľa našich predpokladov získaná pravdepodobne z portálu Trip Advisor. Databáza nám bola pridelená v jednom súbore a to vo formáte .csv. Bližšie popisné informácie neobsahovala, preto sme museli samostatne identifikovať čo pravdepodobne reprezentuje.

Databáza na prvý pohľad obsahuje 20 stĺpcov a 505 riadkov. Tie môžme rozdeliť na 2 základné časti a to hodnotenie a informácie o uživateľovi a popis hotela. Popis a naša interpretácia významu jednotlivých stĺpcov:

Informácie o hodnotiteľovi a hodnotení:

- User country - Krajina z ktorej hodnotiteľ pochádza.
- Number of reviews - Počet hodnotení, ktoré hodnotiteľ napísal na danom portály.
- Number of hotel reviews - Počet hotelov, ktoré hodnotiteľ hodnotil.
- Helpful votes - Počet 'lajkov' od iných užívateľov, ktorí hodnotili toto hodnotenie ako nápomocné.
- Score - Skóre aké hodnotiteľ hotelu pridelil.
- Period of stay - Veľmi široký časový interval v akom hodnotiteľ pobudol v danom hotely.
- User continent - Kontinent z ktorého hodnotiteľ pochádza.
- Member years - Počet rokov, koľko je hodnotiteľ členom portálu Trip Advisor.
- Review month - Mesiac v ktorom bolo hodnotenie pridané.
- Review weekday - Deň v týždni v ktorom bolo hodnotenie pridané.

Informácie o hodnotenom hoteli:

- Pool - Popis hotela, či zahŕňal bazén.
- Gym - Popis hotela, či zahŕňal fitness.
- Tennis court - Popis hotela, či zahŕňal tenisový kurt.
- Spa - Popis hotela, či zahŕňal wellnes.
- Casino - Popis hotela, či zahŕňal kasíno.

- Free internet - Popis hotela, či zahŕňal internet zdarma.
- Hotel Name - Názov hotela, ktorý bol hodnotený.
- Hotel Star - Počet hviezdičiek hotela, ktorý bol hodnotený.
- Number of rooms - Počet izieb v hoteli.

Ako dosť chýbajúce v tomto datasete považujeme finálne, slovné hodnotenia úživateľa pre daný hotel. Dolovanie z textov by sme považovali za najzaújimavejší spôsob získavania informácií z hodnotení na rôzných portáloch, ktoré poskytyjú recenzie a pomáhajú tým dalším uživateľom v rozhodovaní, kde uskutočniť svoj budúci pobyt. Recenzie, ktoré obsahujú len jedno číselné hodnotenie nie su veľmi detailné a objektívne.

3 Predspracovanie dát

Dáta sme importovali do progrmau RapidMiner. Ten nám automaticky odporučil aké dátové typy tieto dáta reprezentujú, či sa jedná o binárné dáta, celé či reálne čisla, alebo textové reťazce.

Názvy stĺpcov obsahovali medzery, bodky, veľké či malé písmená. To považujeme pre budúce strojové spracovanie ako nevhodné. Preto sme jednotlivé názvy v prvom kroku premenovali. A aspoň ich zbavili medzier a bodiek. To všetko pomocou modulu `Rename` programu RapidMiner.

Ďalším krokom bolo skontrolovať dáta pre správne minimálne a maximálne hodnoty v intervaloch a rovnako aj ich správny formát. Pomocou nástroja sme objavili pár nezrovnalostí ako záporné hodnoty v stĺpci, ktoré reprezentovali koľko rokov je daný úživateľ členom portálu. Pomocou modulu `Generate attributes` sme napísali jednoduchý if/then príkaz, ktorý keď v stĺpci `Member year` uvidí záporné čislo, tak ho zmení na číslicu nula.

Všetky binárne hodnoty v našich dátach obsahovali hodnoty `YES` a `NO`, čo sa pre strojové spracovanie nehodí, ak to majú byť data binárne, chceme aby obsahovali hodnoty najlepšie `true` a `false` alebo číslice `0` a `1`. Začo nám RapidMiner určite v dalšom kroku poďakuje. S použitím modulu z predchodzieho kroku sme ľahko tieto dáta zmenili na naozaj binárne hodnoty.

V poslednom kroku spracovania dát sme skontrolovali či RapidMiner správne určil, že čiselné dáta sú `integer` alebo `real` a nezrovnalosti ručne opravili.

4 Prvotné dolovanie dát

V prvých krokoch našej práce s programom RapidMiner, sme si vyskúšali čo zaujímave môžme získať bez použitia nejakých zložitejšich algoritmov a tým sa zoznámili s dátami samotnými.

1. *Aké priemerné honotenie udeľujú užívatelia poďla kontinentu?*
   Pomocou modulu `Set role` sme si určili `Score` ako náš `label` atribút a ďalej programom filtrovali potrebné dáta pomocou modulu `Select Atributes`. V závere pomocou modulu `Agregate` sme mohli použiť všeobecne známu funkciu priemer.
   Takýmto štýlom sme obdobne postupovali aj v dalších úlohách zmienených v tejto kapitole. A to kombinovaním `label` atribútov a filtrovaním dát v závere s použitim `Aggregate` modulu.

Row No.	User_continent	average(Score)
1	Africa	3.429
2	Asia	3.778
3	Europe	4.153
4	North America	4.159
5	Oceania	4.146
6	South America	4.429

2. *Aké priemerné honotenie udeľujú uživateľa podľa dôvodu ich cesty a koľko ich je?*

Row No.	Traveler_type	average(Score) ↓	count(Traveler_type)
4	Friends	4.256	82
2	Couples	4.234	214
3	Families	4.018	110
5	Solo	3.917	24
1	Business	3.878	74

3. *V akom dni v týždni udeľujú užívatelia najpozitívnejšie hodnotenia?*

Row No.	Review_we...	average(Score) ↓	count(Traveler_type)
3	Saturday	4.361	61
1	Friday	4.215	65
5	Thursday	4.161	62
4	Sunday	4.156	77
2	Monday	4.135	74
7	Wednesday	4	85
6	Tuesday	3.925	80

4. *Kto sú hodnotiteľia s najväčším počtom udelených hodnotení pre hotely a aká je táto hodnota?*

Row No.	Nr_reviews	Nr_hotel_reviews ↓	Traveler_type	average(Score)
439	290	263	Business	4
440	290	263	Couples	4
444	415	162	Business	4
422	156	126	Friends	3
447	608	117	Business	5
433	235	111	Friends	5
404	118	107	Business	5
437	275	79	Couples	4
370	79	78	Friends	4
403	116	78	Couples	5
443	372	78	Business	4
434	240	76	Couples	5
436	262	75	Families	4
432	189	72	Solo	4
427	167	71	Couples	4.500

5. *Aký hotel je ten najlepšie hodnotený z pomedzi našich uživateľov a odpovedá to jeho počtu hviezdičiek, aké ma pridelené?*

Row No.	Hotel_name	average(... ↓	Hotel_stars
21	Wynn Las Vegas	4.625	5
14	The Venetian Las Vegas Hotel	4.583	5
4	Encore at wynn Las Vegas	4.542	5
8	Marriott's Grand Chateau	4.542	3.500
13	The Palazzo Resort Hotel Casino	4.375	5
18	Trump International Hotel Las Vegas	4.375	5
20	Wyndham Grand Desert	4.375	3.500
11	The Cosmopolitan Las Vegas	4.250	5
1	Bellagio Las Vegas	4.208	5
19	Tuscany Las Vegas Suites & Casino	4.208	3
7	Hilton Grand Vacations on the Boulevard	4.167	3.500
2	Caesars Palace	4.125	5
12	The Cromwell	4.083	4.500
10	Paris Las Vegas	4.042	4
17	Tropicana Las Vegas - A Double Tree by Hilton Hotel	4.042	4
6	Hilton Grand Vacations at the Flamingo	3.958	3
16	Treasure Island- TI Hotel & Casino	3.958	4
15	The Westin las Vegas Hotel Casino & Spa	3.917	4
5	Excalibur Hotel & Casino	3.708	3
9	Monte Carlo Resort&Casino	3.292	4
3	Circus Circus Hotel & Casino Las Vegas	3.208	3

5 Dolovacia úloha

5.1 Formulácia

Užívatelia z rôznych kútov sveta preferujú rôzne vybavenie hotelov. Zatial čo američania zbožnujú hazard, to nemusí platiť o obyvateľoch Indie, ktorí hotely s kasínom vyhľadávať nemusia, atd. V tejto úlohe sa budeme snažiť zamerať práve na vybavenie hotelov v provnaní s pôvodom ich zákazníkov. Takýmto spôsobom spravíme hotelom prieskum ktoré z ich vybavení hraje úlohu pre ktorú cieľovú skupinu. Či už z pohľadu národností, alebo typu zákazníka.

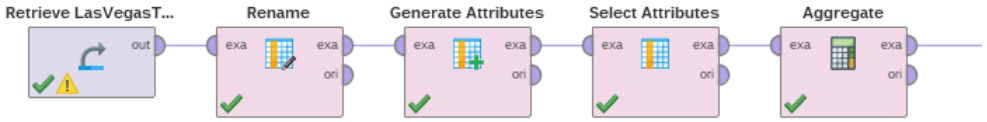
5.2 Postup

V prvom kroku sme sa rozhodli odflitrovať najčastejšie sa vyskytujúce národnosti v našom datasete. Vo výsledku sme zistili, že to sú USA, UK, Kanada, Austrália, Írsko a India. Odfiltrovali sme ich preto, aby nám tie najmenej početné národnosti obsiahnuté v datasete zbytočne neskreslili a nespriehľiadnili výsledok.
Pri príprave dát poďla typu `Traveler type` - Friends, Couples, Families, Solo, Business, nebola filtrácia potrebná, kedže boli zastupené rovnomerne, viď. kapitola 4.

Najčastejšie národnosti v hodnotení

Row No.	User_country	count(Us... ↓
47	USA	217
46	UK	72
4	Canada	65
1	Australia	36
21	Ireland	13
18	India	11

Hotely si pre lepšiu demonštráciu úholy vymenujeme aj spoločne so službami ktoré ponúkajú.

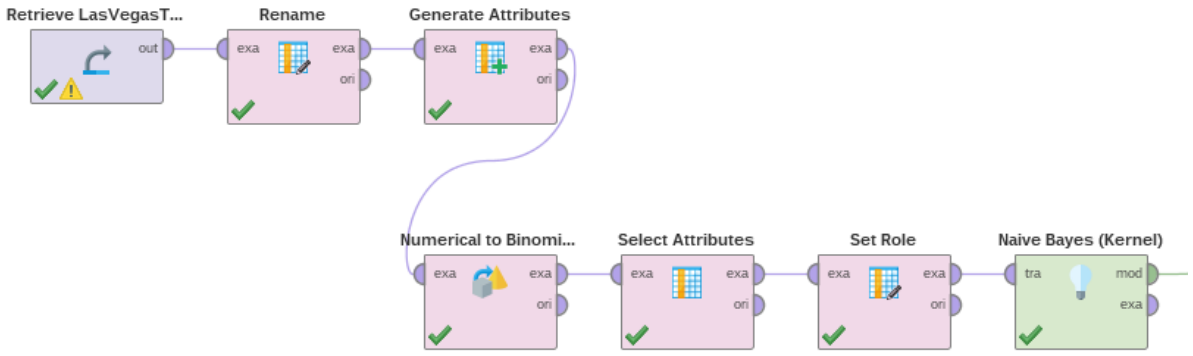


Row No.	Hotel_name	Pool	Free_internet	Casino	Spa	Tennis_court	Gym
1	Bellagio Las ...	1	1	1	1	0	1
2	Caesars Pal...	1	1	1	1	0	1
3	Circus Circus...	0	1	1	0	0	1
4	Encore at wy...	1	1	1	1	0	1
5	Excalibur Hot...	1	1	1	1	0	1
6	Hilton Grand ...	1	1	0	0	0	1
7	Hilton Grand ...	1	1	1	1	0	1
8	Marriott's Gra...	1	1	1	0	0	1
9	Monte Carlo ...	1	0	1	1	0	1
10	Paris Las Ve...	1	1	1	1	0	1
11	The Cosmop...	1	1	1	1	0	1
12	The Cromwell	1	1	1	0	0	0
13	The Palazzo ...	1	1	1	1	0	1
14	The Venetian ...	1	1	1	1	0	1
15	The Westin la...	1	1	1	1	0	1
16	Treasure Isla...	1	1	1	1	1	1
17	Tropicana La...	1	1	1	1	1	1
18	Trump Intern...	1	1	1	1	0	1
19	Tuscany Las ...	1	1	1	1	1	1
20	Wyndham Gr...	1	1	0	0	1	1
21	Wynn Las Ve...	1	1	1	1	1	1

5.3 Výsledky

Vo finalnom výsledku s odfiltrovanými narodnostiami, čí typom zákaznika prevedieme porovnanie pomocou Naive Bayes algoritmu. Tak aby sme v závere dostali grafickú reprezentáciu toho, ako veľmi si daná národnosť či typ zákaznika vybrali svoj hotel. Pre porovnanie sme si zo všetkých služieb vybrali práve kásino. Kde na prvom grafe, obrázok 2, môžme vidieť, že pre obyvateľov Indie nie je kasíno v hoteli

až tak nevyhnutné ako napríklad pre obyvateľov Veľkej Británie. V druhom grafe, obrázok 3 sme rovnaké porovnanie s kasínom previedli na typ zákazníka. Môžme pozorovať, že pre zákaznika, ktorý cestuje sám, je kasíno v hoteli nevyhnutná vec, zatiaľ čo pre rodinu to až tak veľkú rolu nehraje. Dalšie výsledky porovnaní sme do dokumentácie nevkladali, pre ich rozsiahlu veľkost a môžete ich nájst v adresári projektu `output_data/`.



Obr. 1: Fínalne spracovanie výsledku v prostredí RapidMiner.

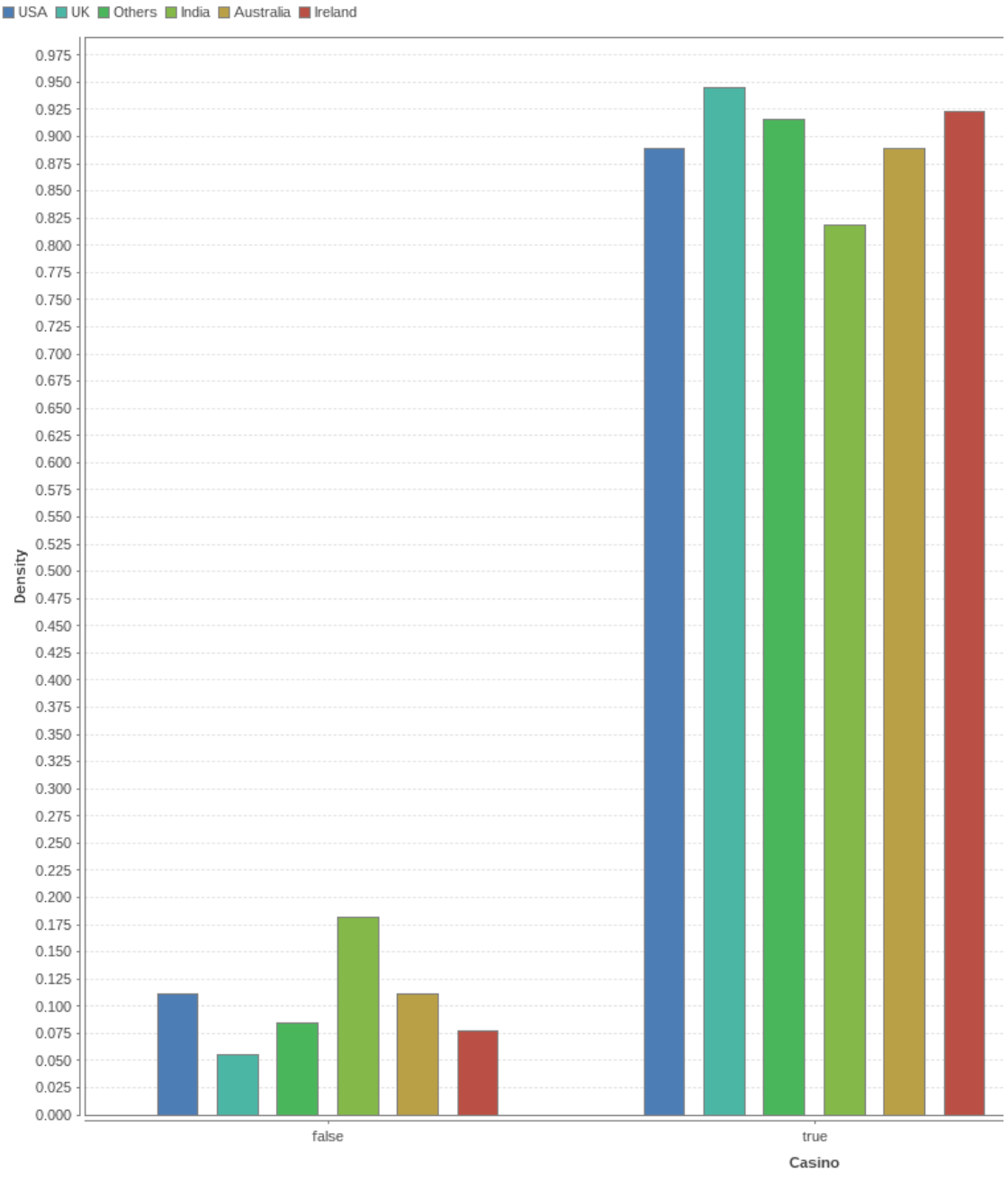
6 Záver

V tomto projekte sme sa venovali spracovaniu veľkých objemov dát a získavaniu znalostí z nich, pomocou programu RapidMiner. Prešli sme si celým procesom od získania, predspracovania, spracovania, analýze a výsledku z dát.

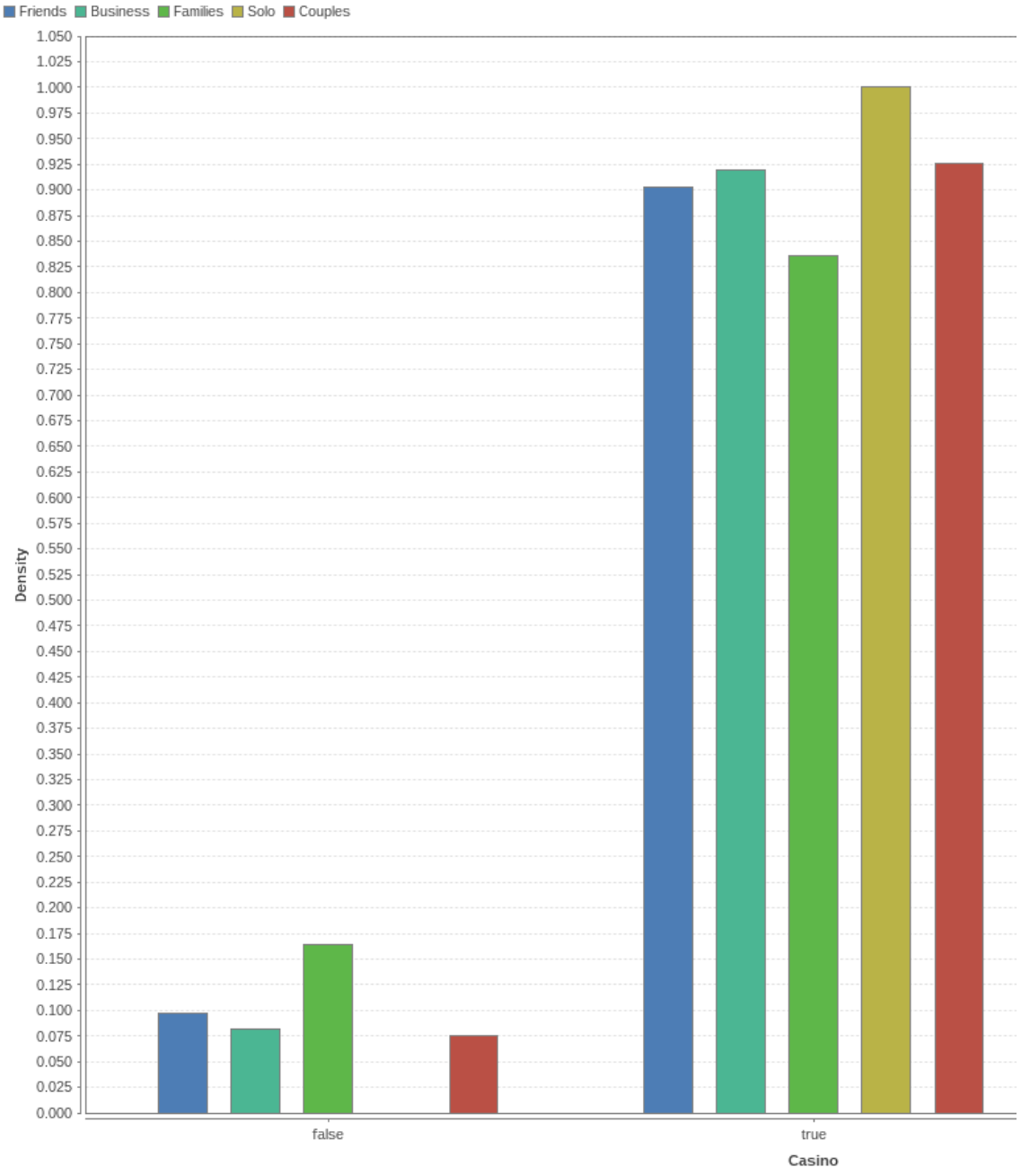
Pracovali sme s databázov hotelov z ktorej sa nám podarilo pomocou dolovacích algoritmov získať zaujímave spojenie služieb, ktoré hoteli ponúkajú v porovnaní z akej krajiny ich zakazníci pochadzajú a aký typ cestovateľa predstavujú.

V budúcnosti by sme odcenili ak by tento data set obsahoval v dátach napríklad slovné hodnotenie či cenu pobytu vďaka čomu by sme vedeli vymyslieť zaujímavejšie zadanie úloh.

Projekt hodnotíme ako zaujímavy a príjemný na spracovanie za čo vďačíme hlavne nástroju RapidMiner s ktorým je dolovanie znalostí veľmi intuitívne a jednoduché.



Obr. 2: Graf reprezentujúci, či si zakazník danej krajiny vybral hotel ktorý kasino obsahoval, alebo nie.



Obr. 3: Graf reprezentujúci, či si zákazník daného typy vybral hotel ktorý kasino obsahoval, alebo nie.