

Analiza podataka iz skupa Bakery

Seminarski rad u okviru kursa
Istraživanje podataka 1
Matematički fakultet

Katarina Rudinac, 8/2015 ^{*}
Dimitrije Špadijer, 398/2016 [†]

28. avgust 2018

Sažetak

U ovom radu predstavljeni su različiti metodi i tehnike koji se koriste u istraživanju podataka. Korišćena je baza podataka o transakcijama jednog lanca pekara. Obradene su oblasti analize i pretprocesiranja podataka, pravila pridruživanja, klaster analize, kao i klasifikacije.

Sadržaj

1	Uvod	1
2	Analiza i priprema podataka	2
2.1	Priprema podataka	2
2.2	Najprodavaniji artikli	4
3	Pravila pridruživanja	6
3.1	Artikli koji se prodaju zajedno	7
3.2	Artikli koji se prodaju zajedno radnim danom i vikendom	7
3.3	Artikli koji se prodaju zajedno na različitim prodajnim mestima	8
4	Klaster analiza	10
4.1	Klasterovanje zaposlenih	10
4.2	Klasterovanje prodajnih mesta	12
5	Klasifikacija	14
6	Zaključak	15

1 Uvod

Baza podataka „Bakery” sadrži podatke o prometu iz dvadeset pekara u jednom lancu iz Sjedinjenih Američkih Država, kao i o zaposlenima na svakoj lokaciji i proizvodima koje pekara nudi. Podaci su preuzeti sa veba na lokaciji <http://poincare.matf.bg.ac.rs/~nenad/ip1/podaci/bakery.7z>. Za istraživanje podataka korišćen je alat KNIME.

^{*}mi15008@alas.matf.bg.ac.rs

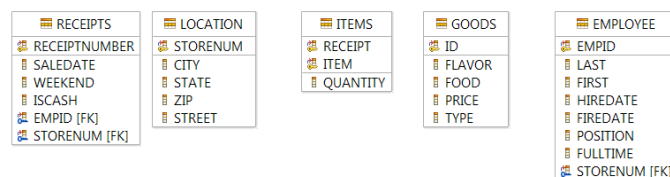
[†]mm11021@alas.matf.bg.ac.rs

2 Analiza i priprema podataka

U ovom poglavlju opisana je struktura podataka u bazi „Bakery” i izdvojene su dve tabele u obliku .csv datoteka, koje se koriste u daljem radu. Podaci su zatim analizirani i izneti su dobijeni zaključci.

2.1 Priprema podataka

Baza podataka „Bakery” sastoji se od 4 entiteta, a to su računi (redni broj, datum izdavanja, da li je u pitanju vikend, da li je plaćen gotovinom, identifikacioni broj zaposlenog koji ga je izdao i redni broj pekare u kojoj je izdat), pekare (redni broj, grad, država, zip kod i ulica), artikli (identifikacioni broj, ukus, vrsta, cena i tip) i zaposleni (redni broj, prezime, ime, datum zapošljavanja, datum otpuštanja, pozicija, da li radi puno radno vreme i redni broj pekare u kojoj radi). Jedan zaposleni radi u jednoj pekari, dok u jednoj pekari može raditi više zaposlenih. Jedan račun može izdati jedan prodavac u jednoj pekari, dok jedan prodavac (u jednoj pekari) može izdati više računa. Jedan račun se može sastojati od više artikala i jedan artikal se može izdati na više računa. Podaci o tome se nalaze u zasebnoj tabeli. Takođe, u toj tabeli se nalazi količina koliko je nekog artikla prodato na jednom računu.



Slika 1: Struktura baze podataka „Bakery”

Zanimljivo je izvući zaključke o tome koji je artikal najprodavaniji i koji se artikli kupuju zajedno. Za to je potrebna tabela u kojoj se nalaze informacije o računima i artiklima. Međutim, tabela „ITEMS” u bazi je (namerno) osiromašena, jer sadrži samo identifikatore, a potrebno je da se u njoj nalaze i informacije o artiklu i o računu. Zato je nad bazom izvršen SQL upit kojim je izdvojena odgovarajuća tabela i njen sadržaj je sačuvan u datoteci „bakery.csv”.

```
select r.*, i.quantity, gd.*
from PEKARA.receipts as r join PEKARA.items as i on r.receiptnumber=i.receipt
      join PEKARA.goods as gd on i.item=gd.id
order by 1;
```

Slika 2: SQL upit pomoću kojeg je iz baze izvučena „bakery.csv” datoteka

Struktura dobijene tabele prikazana je u tabeli 1. Ova tabela predstavlja detalje o svim izvršenim transakcijama i pogodna je za dalje istraživanje i dobijanje željenih informacija.

Zanimljivo je izvući zaključke i o zaposlenima, podeliti ih u određene grupe u zavisnosti od toga na kojoj su poziciji i koliku zaradu donose pekari od prodaje artikala. Takođe, pogodno je slične zaključke doneti i za prodajna mesta. Prethodna tabela nije pogodna za to, pa je zato drugim SQL upitom iz baze podataka izdvojena nova tabela koja je sačuvana u datoteci „bakery2.csv”.

Naziv polja	Opis
RECEIPTNUMBER	Redni broj računa
SALEDATE	Datum izdavanja računa
WEEKEND	Da li je račun izdat tokom vikenda (true/false)
ISCASH	Da li je plaćeno u gotovini (true/false)
EMPID	Identifikacioni broj zaposlenog
STORENUM	Redni broj pekare u lancu
QUANTITY	Količina prodatog artikla na jednom računu
ID	Identifikacioni broj artikla (razlikuje se za isti proizvod ako je ukus drugačiji)
FLAVOR	Ukus
FOOD	Vrsta artikla
PRICE	Cena artikla
TYPE	Tip artikla (food/drink)

Tabela 1: Struktura tabele u „bakery.csv” datoteci

```

with zarade(empid, zarada) as
(select e.empid, sum(i.quantity*gd.price)
 from PEKARA.employee as e join PEKARA.receipts as r on e.empid=r.empid
      join PEKARA.items as i on r.receiptnumber=i.receipt
      join PEKARA.goods as gd on i.item=gd.id
 group by e.empid)
select e.*, z.zarada, lc.*
 from PEKARA.employee as e join zarade as z on e.empid=z.empid
      join PEKARA.location as lc on e.storenum=lc.storenum

```

Slika 3: SQL upit pomoću kojeg je iz baze izvučena „bakery2.csv” datoteka

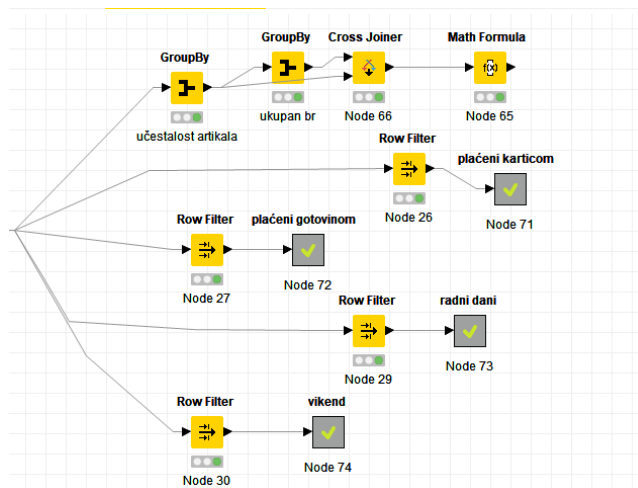
Struktura dobijene tabele prikazana je u tabeli 2. Ona predstavlja informacije o zaposlenima i o njihovim radnim mestima.

Naziv polja	Opis
LAST	Prezime zaposlenog
FIRST	Ime zaposlenog
HIREDATE	Datum zapošljivanja
FIREDATE	Datum otpuštanja
POSITION	Pozicija (manager/shift manager/barista/cashier)
FULLTIME	Da li radi puno radno vreme (true/false)
STORENUM	Redni broj pekare
EMPID	Redni broj zaposlenog
ZARADA	Koliki je ostvaren promet za svakog zaposlenog
CITY	Grad u kojem se nalazi pekara
STATE	Savezna država
ZIP	Zip kod
STREET	Ulica u kojoj se nalazi pekara

Tabela 2: Struktura „bakery2.csv” datoteke

2.2 Najprodavaniji artikli

Od interesa je utvrditi koji artikli su najprodavaniji, kao i da li to zavisi od toga da li su plaćeni kešom ili karticom i da li su kupljeni vikendom ili radnim danima. Artikli su prvo grupisani po računima, zatim su ponovo grupisani kako bi se dobio ukupan broj svih prodanih artikala, onda su spojene tabele kako bi se matematičkom formulom dobila učestalost svakog artikla u ukupno prodanim artiklima.



Slika 4: Korišćeni čvorovi u alatu KNIME

Tabela koja predstavlja artikle poredane po učestalosti prodaje prikazana je na slici 5.

Table "default" - Rows: 18					Spec - Columns: 4	Properties	Flow Variables
Row ID	I S...	S FOOD	I Sum(Q...	D ▼ ne...			
Row0_Row14	53235	Tart	10142	0.191			
Row0_Row1	53235	Cake	8641	0.162			
Row0_Row3	53235	Cookie	7714	0.145			
Row0_Row4	53235	Croissant	4881	0.092			
Row0_Row5	53235	Danish	3086	0.058			
Row0_Row6	53235	Eclair	2937	0.055			
Row0_Row2	53235	Coffee	2826	0.053			
Row0_Row10	53235	Lemonade	2075	0.039			
Row0_Row11	53235	Meringue	1319	0.025			
Row0_Row9	53235	Juice	1318	0.025			
Row0_Row16	53235	Twist	1248	0.023			
Row0_Row12	53235	Pie	1175	0.022			
Row0_Row17	53235	Water	1151	0.022			
Row0_Row8	53235	Frappuccino	1096	0.021			
Row0_Row13	53235	Soda	1025	0.019			
Row0_Row7	53235	Espresso	1008	0.019			
Row0_Row15	53235	Tea	953	0.018			
Row0_Row0	53235	Bear Claw	640	0.012			

Slika 5: Najprodavaniji artikli

Vidi se da su četiri najčešće prodavana artikla tart, torta, keks i kroasan. Postavlja se pitanje da li način plaćanja utiče na to koji su artikli najčešće prodavani. Odgovarajuće tabele se nalaze na slikama 6 i 7.

Row ID	I S...	S FOOD	I Sum(Q...	D ▼ ne...
Row0_Row14	35484	Tart	6797	0.192
Row0_Row1	35484	Cake	5661	0.16
Row0_Row3	35484	Cookie	5260	0.148
Row0_Row4	35484	Croissant	3167	0.089
Row0_Row5	35484	Danish	2003	0.056
Row0_Row6	35484	Eclair	1975	0.056
Row0_Row2	35484	Coffee	1880	0.053
Row0_Row10	35484	Lemonade	1400	0.039
Row0_Row11	35484	Meringue	854	0.024
Row0_Row9	35484	Juice	851	0.024
Row0_Row16	35484	Twist	851	0.024
Row0_Row12	35484	Pie	771	0.022
Row0_Row17	35484	Water	759	0.021
Row0_Row8	35484	Frappuccino	740	0.021
Row0_Row7	35484	Espresso	710	0.02
Row0_Row15	35484	Tea	692	0.02
Row0_Row13	35484	Soda	674	0.019
Row0_Row0	35484	Bear Claw	439	0.012

Slika 6: Najprodavaniji artikli koji su plaćeni karticom

Row ID	I Su...	S FOOD	I Sum(Q...	D ▼ ne...
Row0_Row14	17751	Tart	3345	0.188
Row0_Row1	17751	Cake	2980	0.168
Row0_Row3	17751	Cookie	2454	0.138
Row0_Row4	17751	Croissant	1714	0.097
Row0_Row5	17751	Danish	1083	0.061
Row0_Row6	17751	Eclair	962	0.054
Row0_Row2	17751	Coffee	946	0.053
Row0_Row10	17751	Lemonade	675	0.038
Row0_Row9	17751	Juice	467	0.026
Row0_Row11	17751	Meringue	465	0.026
Row0_Row12	17751	Pie	404	0.023
Row0_Row16	17751	Twist	397	0.022
Row0_Row17	17751	Water	392	0.022
Row0_Row8	17751	Frappuccino	356	0.02
Row0_Row13	17751	Soda	351	0.02
Row0_Row7	17751	Espresso	298	0.017
Row0_Row15	17751	Tea	261	0.015
Row0_Row0	17751	Bear Claw	201	0.011

Slika 7: Najprodavaniji artikli koji su plaćeni gotovinom

Može se uočiti da se učestalost vrlo blago menja u zavisnosti od toga da li je artikal plaćen karticom ili gotovinom i da je redosled najprodavanija četiri artikla ostao potpuno isti. Prema tome, način plaćanja nema veliki uticaj na učestalost prodaje artikala.

Kada se sličnim postupkom utvrđuje zavisnost učestalosti prodaje artikala od toga da li je vikend ili radni dan, redosled druga dva najprodavanija artikla se promeni — radnim danima, nakon tarta je najprodavanija torta, dok je vikendom najprodavaniji keks. Prvi najprodavaniji artikal nije promenio poziciju, kao ni artikli od četvrtog do osmog mesta. Može se zaključiti da su interesovanja kupaca za osam najčešće prodavanih artikala veoma slična, bez obzira na dan u nedelji, s tim što se vikendom više kupuje keks nego torta.

Output data - 2:29:73:67 - Math Formula

File Hilite Navigation View

Table "default" - Rows: 18 Spec - Columns: 4 Properties Flow Variables

Row ID	I Sum(Su...	S FOOD	I Sum(Q...	D ▼ ne...
Row0_Row14	36982	Tart	6927	0.187
Row0_Row1	36982	Cake	6229	0.168
Row0_Row3	36982	Cookie	5240	0.142
Row0_Row4	36982	Croissant	3380	0.091
Row0_Row5	36982	Danish	2161	0.058
Row0_Row6	36982	Eclair	2046	0.055
Row0_Row2	36982	Coffee	2013	0.054
Row0_Row10	36982	Lemonade	1451	0.039
Row0_Row11	36982	Meringue	922	0.025
Row0_Row16	36982	Twist	877	0.024
Row0_Row9	36982	Juice	840	0.023
Row0_Row12	36982	Pie	796	0.022
Row0_Row8	36982	Frappuccino	776	0.021
Row0_Row17	36982	Water	767	0.021
Row0_Row13	36982	Soda	744	0.02
Row0_Row7	36982	Espresso	708	0.019
Row0_Row15	36982	Tea	648	0.018
Row0_Row0	36982	Bear Claw	457	0.012

Slika 8: Najprodavaniji artikli radnim danom

Output data - 2:29:74:67 - Math Formula

File Hilite Navigation View

Table "default" - Rows: 18 Spec - Columns: 4 Properties Flow Variables

Row ID	I Sum(Su...	S FOOD	I Sum(Q...	D ▼ ne...
Row0_Row14	16253	Tart	3215	0.198
Row0_Row3	16253	Cookie	2474	0.152
Row0_Row1	16253	Cake	2412	0.148
Row0_Row4	16253	Croissant	1501	0.092
Row0_Row5	16253	Danish	925	0.057
Row0_Row6	16253	Eclair	891	0.055
Row0_Row2	16253	Coffee	813	0.05
Row0_Row10	16253	Lemonade	624	0.038
Row0_Row9	16253	Juice	478	0.029
Row0_Row11	16253	Meringue	397	0.024
Row0_Row17	16253	Water	384	0.024
Row0_Row12	16253	Pie	379	0.023
Row0_Row16	16253	Twist	371	0.023
Row0_Row8	16253	Frappuccino	320	0.02
Row0_Row15	16253	Tea	305	0.019
Row0_Row7	16253	Espresso	300	0.018
Row0_Row13	16253	Soda	281	0.017
Row0_Row0	16253	Bear Claw	183	0.011

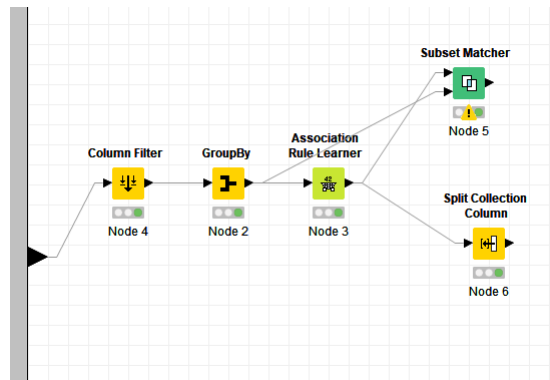
Slika 9: Najprodavaniji artikli vikendom

3 Pravila pridruživanja

U ovom poglavlju tražena je veza između artikala koje pekara prodaje. Od interesa je na osnovu postojećih podataka otkriti koji se artikli najčešće prodaju zajedno. Ako je veliki broj mušterija pored izvesnih artikala kupio još neki, onda se isplati da se ti artikli postave jedni pored drugih, jer postoji velika verovatnoća da će se sledeća mušterija, kada vidi te artikle zajedno, odlučiti da ih kupi sve, iako možda prvobitno to nije želela. Za ovo se koriste pravila pridruživanja.

3.1 Artikli koji se prodaju zajedno

Iz tabele u datoteci „bakery.csv” izdvojene su kolone koje sadrže informacije o broju računa i o vrsti artikla, a zatim su korišćenjem čvora GroupBy svi artikli koji su kupljeni na istom računu spojeni u listu.



Slika 10: Korišćeni čvorovi u alatu KNIME

Korišćen je čvor Association Rule Learner s minimalnom podrškom 0.1 i minimalnom pouzdanošću 0.4. Dobijeno je da je najveća pouzdanost i najbolja lift mera za pravilo da uz dansko pecivo ide tart. Pored toga, tart se s pouzdanošću 0.606 kupuje i uz kroasan. Za oba ova pravila je podrška između 0.1 i 0.2, tj. ovi su artikli kupljeni zajedno u 10–20% svih transakcija.

Row ID	D Support	D Confide...	D Lift	S Conseq...	S implies	[...] Items
rule0	0.142	0.715	1.313	Tart	<---	[Danish]
rule1	0.157	0.414	0.948	Cake	<---	[Cookie]
rule2	0.173	0.456	0.838	Tart	<---	[Cookie]
rule3	0.176	0.606	1.112	Tart	<---	[Croissant]
rule4	0.219	0.403	0.923	Cake	<---	[Tart]
rule5	0.219	0.503	0.923	Tart	<---	[Cake]

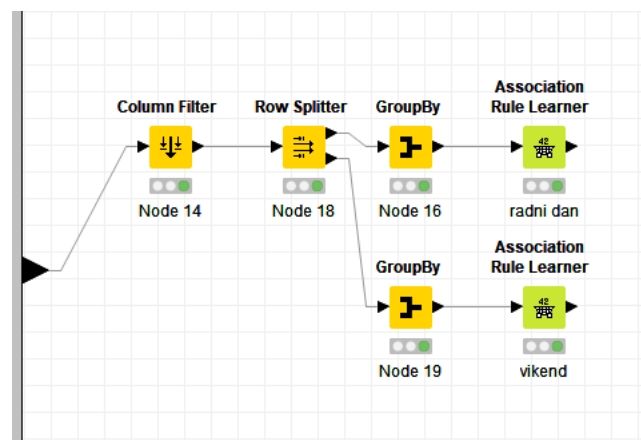
Slika 11: Pravila pridruživanja

Zanimljivo je utvrditi da li se ova pravila ponavljaju nezavisno od prodajnog mesta ili radnog dana. Takođe, postavlja se pitanje da li se mnogo menjaju podrška i pouzdanost.

3.2 Artikli koji se prodaju zajedno radnim danom i vikendom

Da bi se utvrdila pravila pridruživanja samo za transakcije obavljene tokom radnog dana i tokom vikenda, korišćen je čvor Row Splitter kako bi se razdvojile transakcije u zavisnosti od toga kada su obavljene. Postupak razdvajanja se može videti na slici 12.

Za dobijanje novih pravila pridruživanja ponovo je korišćen čvor Association Rule Learner, ali je kao ulaz prosleđena tabela dobijena razdvajanjem. Dobijena pravila pridruživanja prikazana su na slikama 13 i 14.



Slika 12: Razdvajanje transakcija obavljenih radnim danom i vikendom

Row ID	D Support	D Confide...	D Lift	S Conseq...	S implies	[...] Items
rule0	0.135	0.703	1.249	Tart	<---	[Danish]
rule1	0.18	0.596	1.059	Tart	<---	[Croissant]
rule2	0.192	0.486	0.864	Tart	<---	[Cookie]
rule3	0.224	0.544	0.968	Tart	<---	[Cake]

Slika 13: Uočena pravila pridruživanja za transakcije obavljene radnim danom

Row ID	D Support	D Confide...	D Lift	S Conseq...	S implies	[...] Items
rule0	0.144	0.72	1.342	Tart	<---	[Danish]
rule1	0.157	0.423	0.946	Cake	<---	[Cookie]
rule2	0.165	0.443	0.824	Tart	<---	[Cookie]
rule3	0.174	0.611	1.137	Tart	<---	[Croissant]
rule4	0.217	0.405	0.905	Cake	<---	[Tart]
rule5	0.217	0.486	0.905	Tart	<---	[Cake]

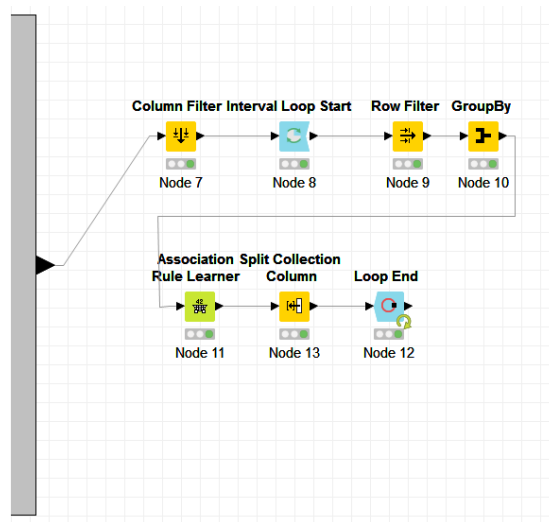
Slika 14: Uočena pravila pridruživanja za transakcije obavljene vikendom

Može se primetiti da, iako je u oba slučaja pravilo sa najvećom pouzdanošću u oba slučaja isto (kupovina danskog peciva povlači kupovinu tarta), pouzdanost, podrška kao i lift mera ponovljenih pravila se blago razlikuju. Razlike između transakcija obavljenih radnim danom i vikendom ogledaju se u tome što se vikendom pojavljuju pravila koja se ne pojavljuju radnim danima — kupovina keksa ili tarta povlači kupovinu torte.

3.3 Artikli koji se prodaju zajedno na različitim prodajnim mestima

Zanimljivo je utvrditi da li su data pravila pridruživanja uopštena za sve pekare ili postoje varijacije u zavisnosti od lokacije na kojoj se pekara nalazi. Koristeći petlju, izdvojena su pravila pridruživanja za svaku od pekara i objedinjena su u jednu krajnu tabelu.

Dobijena pravila pridruživanja se razlikuju među različitim pekarama. Pravilo s najvećom pouzdanošću kada se posmatraju sva prodajna mesta zajedno



Slika 15: Korišćeni čvorovi za razdvajanje transakcija po prodajnim mestima

(uz dansko pecivo se prodaje tart s pouzdanošću 0.715), javlja se u svakoj pekari, a njihove pouzdanosti malo variraju od pekare do pekare (od 0.7 do 0.8), što je slično kao i kada se razmatraju sve pekare zajedno. Lift mera takođe varira (od 1.3 do 1.5).

Row ID	D Support	D Confid...	D Lift	S Conse...	S implies	S Split V...	Iteration
rule0#0	0.122	0.525	1.171	Cake	<---	Danish	0
rule1#0	0.161	0.695	1.358	Tart	<---	Danish	0
rule2#0	0.169	0.448	0.998	Cake	<---	Cookie	0
rule3#0	0.173	0.611	1.194	Tart	<---	Croissant	0
rule4#0	0.22	0.491	0.96	Tart	<---	Cake	0
rule5#0	0.22	0.431	0.96	Cake	<---	Tart	0
rule0#1	0.111	0.583	1.089	Tart	<---	Eclair	1
rule1#1	0.143	0.8	1.493	Tart	<---	Danish	1
rule2#1	0.159	0.421	0.786	Tart	<---	Cookie	1
rule3#1	0.187	0.495	1.048	Cake	<---	Cookie	1
rule4#1	0.198	0.649	1.212	Tart	<---	Croissant	1
rule5#1	0.202	0.429	0.8	Tart	<---	Cake	1
rule0#2	0.155	0.406	0.777	Tart	<---	Cookie	2
rule1#2	0.167	0.733	1.403	Tart	<---	Danish	2
rule2#2	0.174	0.455	1.037	Cake	<---	Cookie	2
rule3#2	0.182	0.565	1.08	Tart	<---	Croissant	2
rule4#2	0.212	0.483	0.924	Tart	<---	Cake	2
rule5#2	0.212	0.406	0.924	Cake	<---	Tart	2
rule0#3	0.106	0.634	1.368	Cake	<---	Danish	3
rule1#3	0.11	0.551	0.955	Tart	<---	Coffee	3
rule2#3	0.13	0.78	1.352	Tart	<---	Danish	3
rule3#3	0.146	0.667	1.155	Tart	<---	Croissant	3
rule4#3	0.191	0.54	0.936	Tart	<---	Cookie	3
rule5#3	0.252	0.544	0.942	Tart	<---	Cake	3
rule6#3	0.252	0.437	0.942	Cake	<---	Tart	3
rule0#4	0.112	0.519	1.083	Cake	<---	Coffee	4
rule1#4	0.112	0.435	0.909	Cake	<---	Croissant	4
rule2#4	0.132	0.8	1.413	Tart	<---	Danish	4
rule3#4	0.174	0.677	1.197	Tart	<---	Croissant	4
rule4#4	0.174	0.442	0.781	Tart	<---	Cookie	4
rule5#4	0.19	0.484	1.01	Cake	<---	Cookie	4
rule6#4	0.236	0.416	0.868	Cake	<---	Tart	4
rule7#4	0.236	0.491	0.868	Tart	<---	Cake	4

Slika 16: Pravila pridruživanja na različitim prodajnim mestima

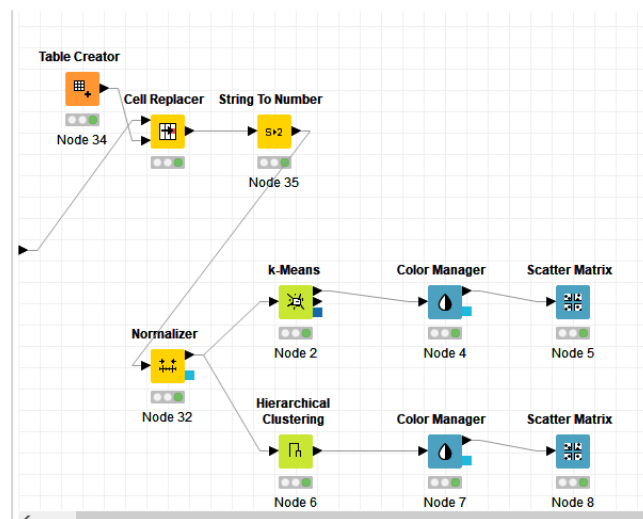
Uočavaju se još neka od pravila pridruživanja kada se razmatraju sva prodajna zajedno (uz keks se prodaje torta ili tart), ali se ona na nekim prodajnim mestima ne pojavljuju (ili se pojavljuju s pouzdanošću manjom od minimalne zadate pouzdanosti). S druge strane, na određenim prodajnim mestima pojavljuju se nova pravila pridruživanja (s pouzdanošću većom od minimalne zadate), kao što su da se uz ekler prodaje tart na prodajnom mestu s identifikatorom 1 s pouzdanošću 0.583 i lift merom 1.089, ili da se uz kafu prodaje torta na prodajnom mestu s identifikatorom 4 s pouzdanošću 0.519 i lift merom 1.083.

4 Klaster analiza

U ovom poglavlju prikazane su tehnike klaster analize. Podaci se dele u određene grupe (klaster) takve da se u okviru iste grupe nalaze slični objekti, a u okviru različitih grupa se nalaze manje slični objekti. Korišćeni su algoritmi K-sredina i hijerarhijskog klasterovanja.

4.1 Klasterovanje zaposlenih

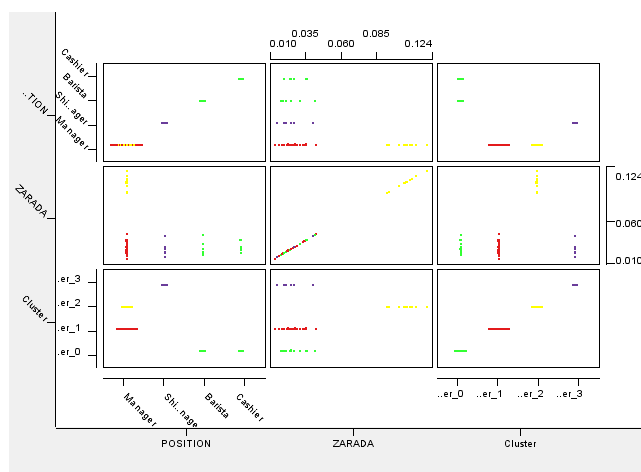
Za primenu algoritama klaster analize korišćena je druga datoteka sa podacima, „bakery2.csv”, u kojoj se nalaza podaci o zaposlenima, među kojima su najzanimljiviji podaci o tome na kojoj poziciji radi zaposleni, kao i koliki je promet ostvario tokom rada u pekari. U poglavlju 2 može se videti kako je skup u „bakery2.csv” dobijen iz baze podataka, kao i atributi koji postoje u skupu (tabela 2). S obzirom da je pozicija kategorički atribut i samim tim nije pogodna za klaster analizu, bilo je potrebno uraditi još neke pripreme podataka. Dodata je nova kolona u kojoj je svakoj poziciji pridružena vrednost - menadžeru 5, menadžeru smene 4, baristi 2 i osobi za kasom 1. Takođe je izvršena normalizacija novih vrednosti i prometa po zaposlenom decimalnim skaliranjem.



Slika 17: Korišćeni čvorovi u alatu KNIME

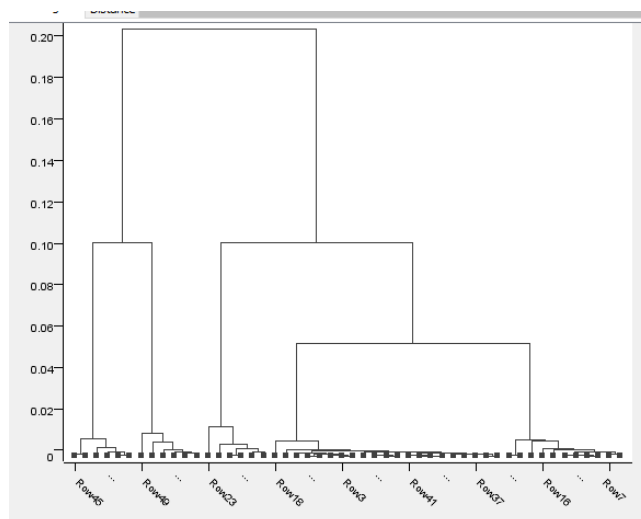
Algoritmom K-sredina, na osnovu pozicije zaposlenog i ostvarenog prometa, dobijena su četiri klastera zaposlenih. Može se primetiti da su menadžeri po-

deljeni u dva klastera zbog značajne razlike u prometu između menadžera, da su menadžeri smene u trećem klasteru, a da bariste i osobe za kasom zajedno formiraju poslednji klaster. Što se prometa tiče, zaposleni koji su ostvarili veliki promet pripadaju jednom klasteru, dok su ostali raspoređeni u preostala tri klastera.



Slika 18: Grafički prikaz dobijenih klastera

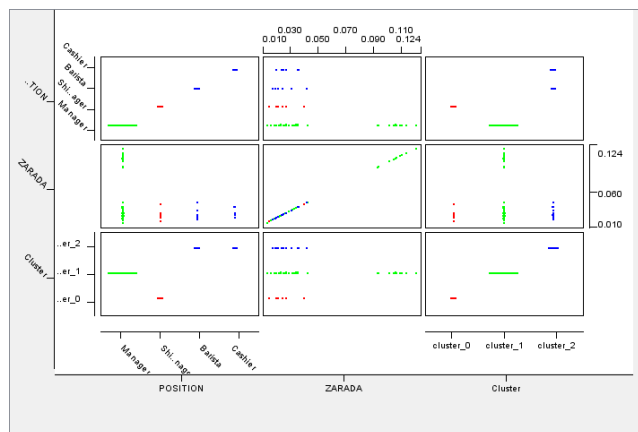
Nakon toga, primenjen je algoritam hijerarhijskog klasterovanja, na osnovu istih kriterijuma kao malopre. Na slici 19 je prikazan dendrogram hijerarhijskog klasterovanja.



Slika 19: Dendrogram hijerarhijskog klasterovanja

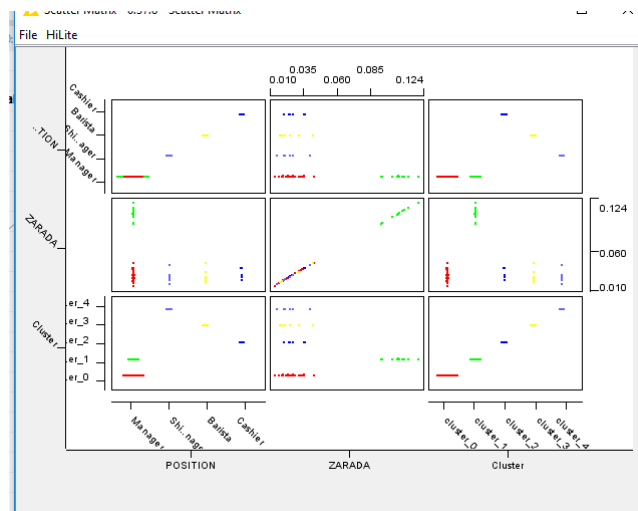
Na dendrogramu se može primetiti da je prirodno posmatrati četiri ili pet klastera. Više od pet klastera nije prirodno, jer bi dosta sličnih podataka veštački otišlo u različite klustere. Ukoliko se iz hijerarhije izdvoje tri klastera, menadžeri

ostaju u istom klasteru bez obzira na ostvaren promet, dok bariste i osobe za kasom ostaju u istom klasteru. Dakle, tri klastera nisu dovoljna.



Slika 20: Grafički prikaz dobijenih klastera kada ih ima 3

Algoritmom K-sredina sa postavljenim brojem klastera na pet, osobe za kasom i bariste se razdvajaju a ostali klasteri ostaju nepromenjeni. Ovo potvrđuje zaključak da je prirodno posmatrati četiri ili pet klastera.

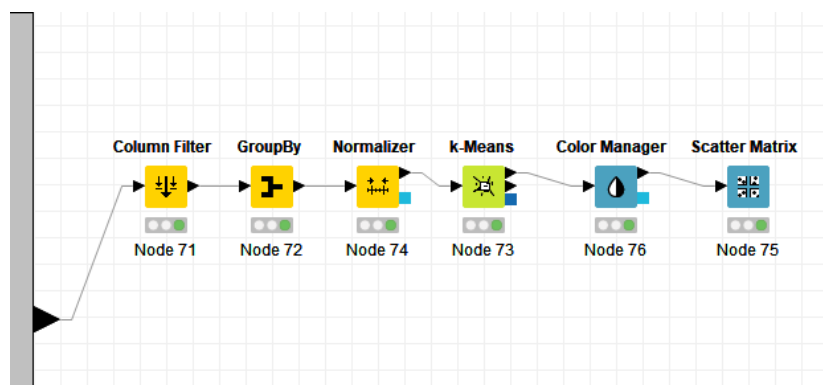


Slika 21: Grafički prikaz dobijenih klastera kada ih ima 5

4.2 Klasterovanje prodajnih mesta

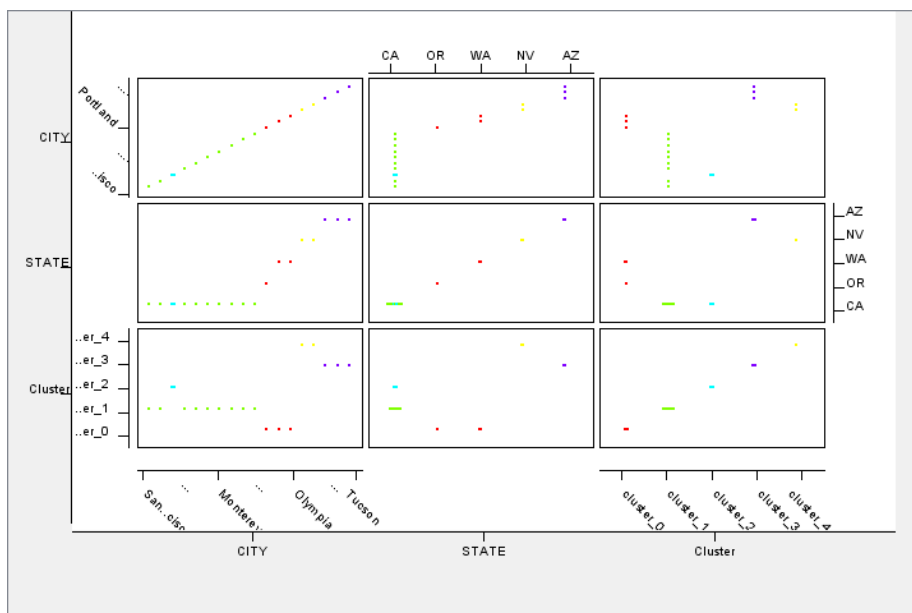
Sličan postupak se može primeniti kako bi se klasterovalo dvadeset prodajnih mesta. Prvo je potrebno eliminisati kolone koje se odnose na zaposlene kako bi ostale samo one koje se odnose na prodajna mesta. Zatim je potrebno grupisati ostvarene zarade na osnovu rednog broja prodajnog mesta. Nakon toga je potrebno normalizovati podatke koji će biti korišćeni u klasifikaciji, zip

kod i zaradu, koristeći decimalno skaliranje. Umesto razdvajanja kategoričkih atributa „CITY” i „STATE” na binarne attribute koji ukazuju na to kom gradu, odnosno državi, pripada prodajno mesto, odlučeno je da se iskoristi zip kod, jer donekle čuva informaciju o tome koliko su mesta blizu jedno drugom.



Slika 22: Korišćeni čvorovi

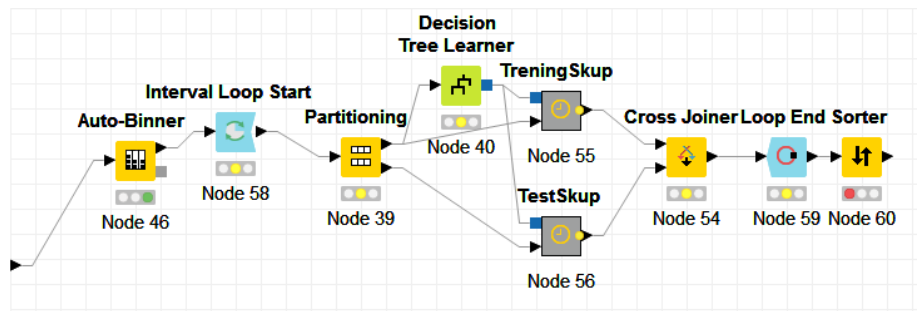
Rezultati nakon primene algoritma K-sredina su pokazali da zarada nije uticala na pripadnost klasteru i da su prodajna mesta iz istih saveznih država uglavnom grupisane u iste klaster. Međutim, u okviru prodajnih mesta u Kaliforniji, one iz Los Anđelesa su izdvojene u zaseban klaster, dok je prodajno mesto u Oregonu stavljena u isti klaster sa prodajnim mestima iz države Vašington.



Slika 23: Grafički prikaz dobijenih klastera

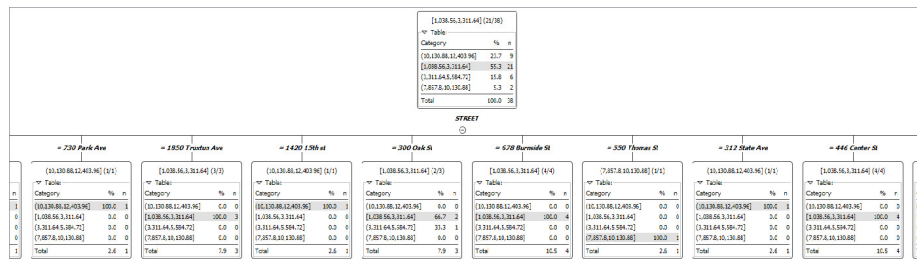
5 Klasifikacija

Svako od zaposlenih je tokom svog rada naplatio od mušterija određenu količinu novca i time doprineo ukupnoj zaradi lanca pekara. Neko je ostvario veću, a neko manju zaradu. Od interesa je podeliti interval od najmanje do najveće ostvarene zarade na određen broj manjih intervala i klasifikovati zaposlene prema tome u kom intervalu se nalazi njihova zarada. Ovakva klasifikacija je korisna, jer ukoliko vlasnik lanca pekara želi zaposliti novog zaposlenog na nekoj lokaciji, potrebno je da ima predstavu koliku će mu zaradu taj zaposleni donositi i shodno tome može odlučiti koliko će plaćati novog zaposlenog.



Slika 24: Korišćeni čvorovi u alatu KNIME

Za klasifikaciju je korišćena tehnika klasifikacije koja koristi drvo odlučivanja. Interval ostvarenih zarada podeljen je na 5 podintervala i oni su odabrani kao klase. Za to je korišćen čvor Auto-Binner. Nakon toga su podaci podeljeni na 2 dela, skup za treniranje i skup za testiranje, korišćenjem čvora Partitioning. Njegov izlaz koji predstavlja skup za treniranje dodeljen je kao ulaz za čvor Decision Tree Learner koji treba da na osnovu datog skupa napravi drvo odlučivanja. Dobijeno drvo odlučivanja izgleda ovako:

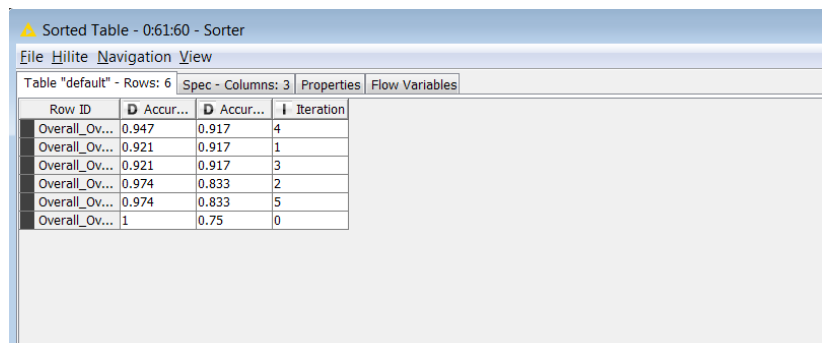


Slika 25: Drvo odlučivanja

Ispostavlja se da se odluka o tome u koju klasu treba svrstati prodavca zavisi od adrese pekare. Takva odluka je smisljena, jer se pekare mogu nalaziti u različitim gradovima ili u različitim delovima istog grada i u nekima će se prodavati više artikala nego u nekim drugim.

Nakon toga je proverena preciznost ovoga modela na skupu za treniranje i zasebno na skupu za testiranje, a potom su ti podaci objedinjeni pomoću čvora Cross Joiner. Rezultati zavise najviše od toga kako se podele podaci, pa je pomoću petlje pre čvora Partitioning omogućeno da se čitav postupak odradi

više puta. Korišćeni su čvorovi Interval Loop Start i Loop End, i dobijeni rezultati su sortirani.



Sorted Table - 0:61:60 - Sorter

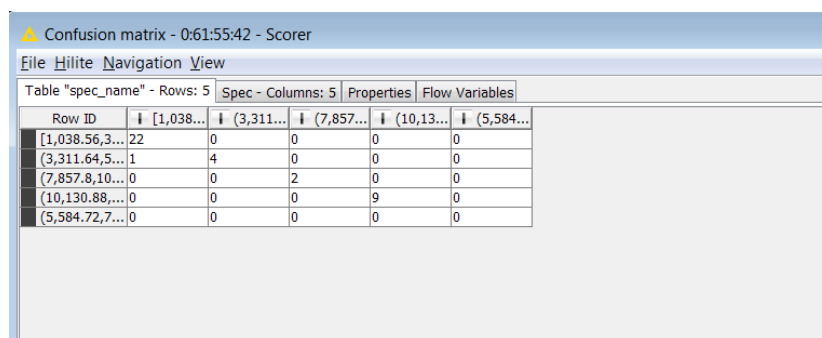
File Hilitte Navigation View

Table "default" - Rows: 6 Spec - Columns: 3 Properties Flow Variables

Row ID	D Accur...	D Accur...	Iteration
Overall_Ov...	0.947	0.917	4
Overall_Ov...	0.921	0.917	1
Overall_Ov...	0.921	0.917	3
Overall_Ov...	0.974	0.833	2
Overall_Ov...	0.974	0.833	5
Overall_Ov...	1	0.75	0

Slika 26: Preciznost modela klasifikacije

Prva kolona je preciznost na skupu za trening, a druga je preciznost na skupu za test. Zaključak je da se klasifikacija vrši s velikom preciznošću, a to se može videti i iz matrica konfuzije za skupove za treniranje i za testiranje.



Confusion matrix - 0:61:55:42 - Scorer

File Hilitte Navigation View

Table "spec_name" - Rows: 5 Spec - Columns: 5 Properties Flow Variables

Row ID	[1,038...	[3,311...	[7,857...	[10,13...	[5,584...
[1,038.56,3...	22	0	0	0	0
[3,311.64,5...	1	4	0	0	0
[7,857.8,10...	0	0	2	0	0
[10,130.88,...	0	0	0	9	0
[5,584.72,7...	0	0	0	0	0

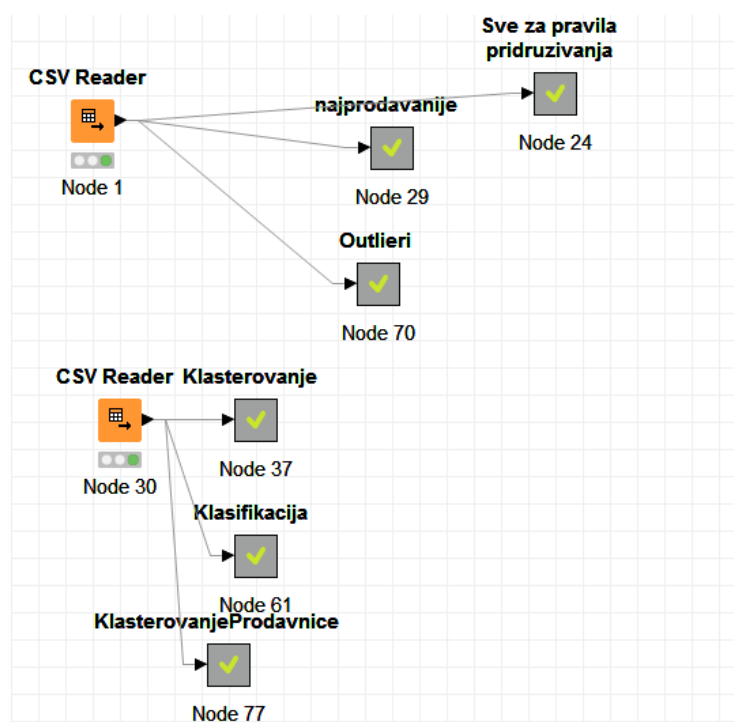
Slika 27: Matrica konfuzije za skup podataka za testiranje

Sama priroda podataka u ovoj bazi je takva da je teško naći mnogo primera za klasifikaciju. Stoga prikazan samo jedan takav primer.

6 Zaključak

U ovom radu predstavljani su različiti metodi i tehnike koji se koriste u istraživanju podataka. Korišćena je baza podataka o transakcijama u okviru jednog lanca pekara u Sjedinjenim Američkim Državama. Prikazano je pripremanje i analiziranje podataka, pravila pridruživanja, klaster analiza i klasifikacija drvetom odlučivanja. Svaki od ovih metoda detaljno je objašnjen i izvedeni su zaključci koji mogu biti korisni vlasniku ovog lanca pekara kako bi unapredio svoje poslovanje.

Za primenu prethodno navedenih metoda i tehnika istraživanja podataka korišćen je alat KNIME. Kompletan prikaz korišćenih čvorova može se videti na slici 28.



Slika 28: Svi korišćeni čvorovi u alatu KNIME