

Saobraćajne nesreće u Francuskoj od 2005. do 2016. godine

Jovan Ležaja

473/2018

Matematički fakultet, Beograd

navoj96@gmail.com

Aleksandar Vračarević

434/2016

Matematički fakultet, Beograd

vracarevicaleksandar@gmail.com

September 2, 2019

1 Uvod

Ovaj rad se fokusira na analizu skupa podataka o saobraćajnim nesrećama u Francuskoj od 2005. do 2006. godine. Pozabavićemo se opisom, analizom i pretprocesiranjem datih podataka, a potom ćemo različitim algoritmima pokušati da pronadjemo pravila pridruživanja (eng. *Association rules*) koristeći se alatima koje nudi IBM SPSS Modeler.

2 Opis podataka

Podaci su preuzeti sa <https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016> i predstavljaju podatke o saobraćajnim nesrećama u Francuskoj prikupljene u period od 2005. do 2016. godine. Kako bismo uopšte pristupili istraživanju skrivenih pravila u okviru ovog skupa, najpre se moramo upoznati sa istim. Naime, skup se sastoji od 5 tabela u .csv formatu. U nastavku ćemo opisati attribute svake od njih.

- `characteristics.csv`
 - **Num_Acc** : identifikator nesreće - numerički
 - **jour** : dan u mesecu - numerički [1-31]
 - **mois** : mesec - numerički [1-12]
 - **an** : poslednje dve cifre godine - numerički [5-16]
 - **hrmn** : vreme u formatu (ssmm) - numerički [1-2.36k]

- **lum** : osvetljenje u trenutku nesreće brojevi [1-5] kodirani na sledeći način:
 - * 1 - dan
 - * 2 - sumrak/zora
 - * 3 - noć bez prisutnog javnog osvetljenja
 - * 4 - noć sa isključenim javnim osvetljenjem
 - * 5 - noć sa uključenim javnim osvetljenjem
- **dep** : INSEE kod odeljenja praćen nulom
- **com** : kod opštine izdat od strane INSEE
- **agg** :
 - * 1 - izvan gradske sredine
 - * 2 - unutar gradske sredine
- **int** : tip raskrsnice [1-9] kodirani na sledeći način:
 - * 1 - van raskrsnice
 - * 2 - X raskrsnica
 - * 3 - T raskrsnica
 - * 4 - Y raskrsnica
 - * 5 - raskrsnica sa više od 4 kraka
 - * 6 - kružni tok
 - * 7 - place
 - * 8 - pružni prelaz
 - * 9 - ostalo
- **atm** : atmosferski uslovi [1-9] kodirani na sledeći način:
 - * 1 - normalni
 - * 2 - slaba kiša
 - * 3 - jaka kiša
 - * 4 - sneg/grâd
 - * 5 - magla/dim
 - * 6 - jak vetar/oluja
 - * 7 - zaslepljujuće vreme
 - * 8 - oblačno
 - * 9 - ostalo
- **col** : tip sudara [1-7] kodiran na sledeći način:
 - * 1 - čeon sudar
 - * 2 - sudar otpozadi
 - * 3 - sudar sa strane
 - * 4 - lančani sudar
 - * 5 - višestruki sudari (više vozila i više sudara)
 - * 6 - drugi sudari
 - * 7 - nesreća bez sudara
- **adr** : poštanska adresa - niska (popunjava se samo za gradske sredine)

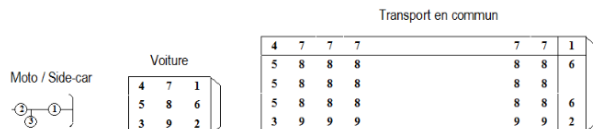
- **gps** : GPS kod - jedan karakter:
 - * M - Métropole
 - * A - Antilles (Martinique or Guadeloupe)
 - * G = Guyane
 - * R = Réunion
 - * Y = Mayotte
- **lat** : geografska širina izražena u broju stepeni
- **long** : geografska dužina izražena u broju stepeni
- **holidays.csv**
 - **ds** : datum nesreće u formatu godina-mesec-dan
 - **holiday** : naziv praznika
- **places.csv**
 - **Num_Acc** : identifikator nesreće - numerički
 - **catr** : kategorija puta [1-9] kodirani na sledeći način:
 - * 1 - autoput
 - * 2 - državni put
 - * 3 - departmentalni putevi
 - * 4 - komunalni putevi
 - * 5 - mreža puteva zabranjena za javnost
 - * 6 - javni parking
 - * 9 - ostalo
 - **voie** : broj puta - numerički
 - **V1** : numerički indeks broja puta (na primer: 2 bis, 3 ter itd.)
 - **V2** : alfanumerički indeks puta
 - **circ** : tip saobraćanja [1-4] kodiran na sledeći način:
 - * 1 - jednosmerna ulica
 - * 2 - dvosmerna ulica
 - * 3 - razdvojen kolovoz
 - * 4 -
 - **nbv** : ukupan broj traka na putu - numerički
 - **vosp** : indikator postojanja rezervisane trake [1-3], nezavisno od toga da li se nesreća dogodila u toj traci, kodiran na sledeći način:
 - * 1 - bickilistička traka
 - * 2 - parking za bicikle
 - * 3 - rezervisan kanal
 - **prof** : kategorije puta [1-4] zavisno od nagiba puta, kodirane na sledeći:
 - * 1 - “dish”
 - * 2 - nizbrdica

- * 3 - vrh brda
- * 4 - dno brda
- **pr** : PR broj kuće - numerička vrednost
- **pr1** : udaljenost od najbližeg PR broja izražena u metrima - numerička vrednost
- **plan** : izgled puta na mapi [1-4], kodirano na sledeći način:
 - * 1 - prav put
 - * 2 - zakrivljen ulevo
 - * 3 - zakrivljen udesno
 - * 4 - “S” oblika
- **lartpc** : širina ostrva na ulici, ako postoji - niska
- **larrout** : širina puta namenjena za saobraćaj - niska
- **surf** : stanje terena [1-9], kodiran na sledeći način:
 - * 1 - normalan
 - * 2 - vlažan
 - * 3 - teren
 - * 4 - potopljen
 - * 5 - sneg na terenu
 - * 6 - blatnjav
 - * 7 - poledica na terenu
 - * 8 - masan/zauljen teren
 - * 9 - ostalo
- **infra** : infrastruktura puteva [1-7], kodirana na sledeći način:
 - * 1 - podzemni tunel
 - * 2 - most/nadvožnjak
 - * 3 - uključenje
 - * 4 - pruga
 - * 5 - “carrefour arranged”
 - * 6 - pešačka zona
 - * 7 - ostalo
- **situ** : pozicija nesreće [1-5], kodirana na sledeći način:
 - * 1 - na putu
 - * 2 - u zaustavnoj traci
 - * 3 - na ivičnjaku
 - * 4 - na trotoaru
 - * 5 - na biciklističkoj stazi
- **env1** : locirano blizu škole - numerička vrednost

● **users.csv**

- **Acc_number** : identifikator nesreće - numerički
- **Num_Veh** : identifikator vozila - alfanumerički

- **place** : pozicija osobe u vozilu u vreme nesreće, kodirano u skladu sa sledećom slikom:



- **catu** : uloga osobe u saobraćaju u trenutku nesreće [1-4], kodirano na sledeći način:
 - * 1 - vozač
 - * 2 - putnik
 - * 3 - pešak
 - * 4 - pešak na rolerima ili skuteru
- **grav** : ozbiljnost povrede [1-4], kodirana na sledeći način:
 - * 1 - neozledjen
 - * 2 - ubijen
 - * 3 - hospitalizovan
 - * 4 - blaga ozleda
- **sex** : pol osobe:
 - * 1 - muško
 - * 2 - žensko
- **Year.on** : godina rođenja - numerički
- **trip** : razlog putovanja [1-9], kodiran na sledeći način:
 - * 1 - kuća-posao
 - * 2 - posao-kuća
 - * 3 - kupovina
 - * 4 - poslovni put
 - * 5 - razonoda
 - * 9 - ostalo
- **secu** : niska koja se sastoji od 2 broja. Prvi označava postojanje sigurnosne opreme [1-9], kodirano na sledeći način:
 - * 1 - pojas za vezivanje
 - * 2 - kaciga
 - * 3 - sedeljka za decu
 - * 4 - reflektujuća oprema
 - * 9 - ostalo

Drugi označava korišćenje sigurnosne opreme [1-3], kodirano na sledeći način:

 - * 1 - oprema je korišćena
 - * 2 - oprema nije korišćena
 - * 3 - neodređeno
- **locp** : pozicija pešaka [1-8], kodirano na sledeći način:

- * 1 - više od 50 metara od pešačkog prelaza
 - * 2 - manje od 50 metara od pešačkog prelaza
 - * 3 - na pešačkom prelazu sa semaforom
 - * 4 - na pešačkom prelazu bez semafora
 - * 5 - na trotoaru
 - * 6 - na ivičnjaku
 - * 7 - pod zaklonom
 - * 8 - u prolazu
 - **actp** : akcija pešaka [0-9], kodirano na sledeći način:
 - * 0 - neodređeno
 - * 1 - kreće se u istom smeru kao i vozilo sa kojim se dogodio sudar
 - * 2 - kreće se u suprotnom smeru kao i vozilo sa kojim se dogodio sudar
 - * 3 - prelazak ulice
 - * 4 - zaklonjen
 - * 5 - u trku
 - * 6 - sa životinjom
 - * 9 - ostalo
 - **etatp** : kategorička vrednost koja određuje da li je pešak bio u društvu drugih ljudi ili ne, kodirano na sledeći način:
 - * 1 - sam
 - * 2 - sa saputnikom
 - * 3 - u grupi ljudi
- **vehicles.csv**
- **Num_Acc** : identifikator nesreće - numerički
 - **Num_veh** : identifikator vozila - alfanumerički kod
 - **GP** :
 - **CATV** : kategorija vozila [01 - 13]
 - * 01 - bicikl
 - * 02 - moped ; 50 kubika
 - * 03 - kvadricikl sa motorom
 - * 04 - suvišno od 2006. (registrovani skuter)
 - * 05 - suvišno od 2006. (motocikl)
 - * 06 - suvišno od 2006. (putnička prikolica za motocikl)
 - * 07 - VL
 - * 08 - neupotrebljena kategorija (VL i karavan)
 - * 09 - neupotrebljena kategorija (VL i prikolica)
 - * 10 - VU
 - * 11 - najviše korišćeno posle 2006. godine (VU(10) + karavan)
 - * 12 - najviše korišćeno posle 2006. godine (VU(10) + prikolica)
 - * 13 - PL samo 3.5T
 - * 14 -

3 Analiza i pretprocesiranje podataka

Prilikom učitavanja tabele *characteristics* smo uočili da je usled loše formatirane datoteke došlo do pogrešne reprezentacije podataka, što smo razrešili jednostavnom *Python* skriptom. Analizirajući tabelu *characteristics* uočili smo da atributi *gps*, *lat* i *long* imaju značajan broj nedostajućih vrednosti (preko 50%), a s obzirom da zamena nekom konkretnom vrednošću nema smisla zato što nemamo dovoljno validnih vrednosti u koloni da njihova zamena bude smisljena, odlučili smo da ih uklonimo, jer smatramo da nam nisu bitni za dalju analizu. Kada je reč o atributima *atm* i *col*, zbog izuzetno malog broja nedostajućih vrednosti (atributi su bili kompletni blizu 100%), u čvoru *Type* smo ih odbacili, jer ne gubimo ništa odbacivanjem tako malog broja podataka. U tabeli se isto tako nalaze i atributi vezani za lokaciju nesreća (ulica, opština, itd.), ali dodatnim posmatranjem smo primetili da je format zapisa tih podataka dosta nekonzistentan, tako da je njihova korisnost dovedena u pitanje, pošto bez iscrpnog analiziranja teksta ne bismo mogli da izvučemo korisne informacije, što je dovelo do odluke da preko čvora *Type* tim atributima postavimo ulogu (‘‘Role’’) na vrednost *None*.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Num_Acc	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
an	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
mois	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
jour	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
hrmm	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
lum	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
agg	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0
int	Continuous	20242	0 None	Never	Fixed		100	839985	0	0	0	0
atm	Continuous	41494	0 None	Never	Fixed		99.993	839930	55	0	0	0
col	Continuous	0	0 None	Never	Fixed		99.999	839974	11	0	0	0
com	Continuous	4990	0 None	Never	Fixed		100	839983	2	0	0	0
adr	Categorical	--	--	Never	Fixed		83.268	699438	0	140547	140547	0
gps	Categorical	--	--	Never	Fixed		43.599	366226	0	473759	473759	0
lat	Continuous	0	0 None	Never	Fixed		43.152	362471	477514	0	0	0
long	Continuous	0	2 None	Never	Fixed		42.77	359258	480727	0	0	0
dep	Continuous	0	0 None	Never	Fixed		100	839985	0	0	0	0

Figure 1: Sadržaj *Data Audit* čvora za tabelu *characteristics*

Analizom skupa podataka *users* uočili smo da atributi *locp*, *actp* i *etatp*, koji predstavljaju informacije vezane za pešaka, imaju značajan broj neodređenih vrednosti (preko 50%), tako da smo te kolone izbacili iz skupa podataka *users*. Kada je u pitanju atribut *secu*, čije su vrednosti predstavljene kao dva broja, uočili smo nekonzistentnost odredenih polja sa zadatim opisom reprezentacije tog atributa, tako da smo te nekonzistentne vrednosti preimenovali u *NA* (neodređenu vrednost). Za svaki atribut koji je imao 0 kao vrednost, a nije bilo definisano šta ta vrednost predstavlja, 0 je zamenjena sa *NA*. Za atribut *place* smo sve vrednosti ostavili kakve jesu, pošto je šema koja predstavlja kodiranje bila nedovoljno jasna.

Skup podataka *vehicles* smo analizirali i zaključili da sve slogove koji sadrže nedostajuće i neodređene vrednosti možemo da odbacimo. Nismo naišli ni na kakve nepravilnosti koje iziskuju detaljnije procesiranje.

Nakon što smo uvideli da kolone *v1* i *v2* skupa *places* sadrže ogroman broj nedostajućih vrednosti, odbacili smo ih. Pošto se u kolonama *pr* i *pr1* javlja preko 50% nedostajućih vrednosti, a smatramo da ne postoji smislen način da te

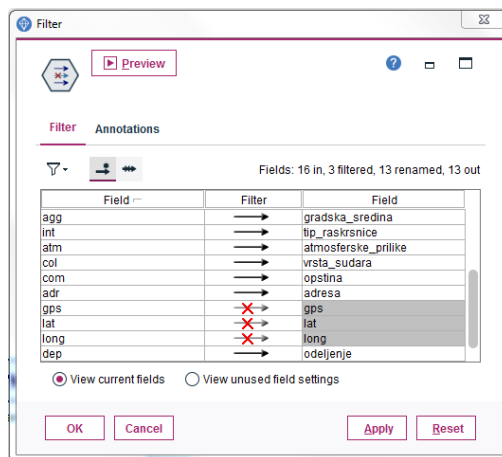


Figure 2: Sadržaj *Filter* čvora za tabelu **characteristics**

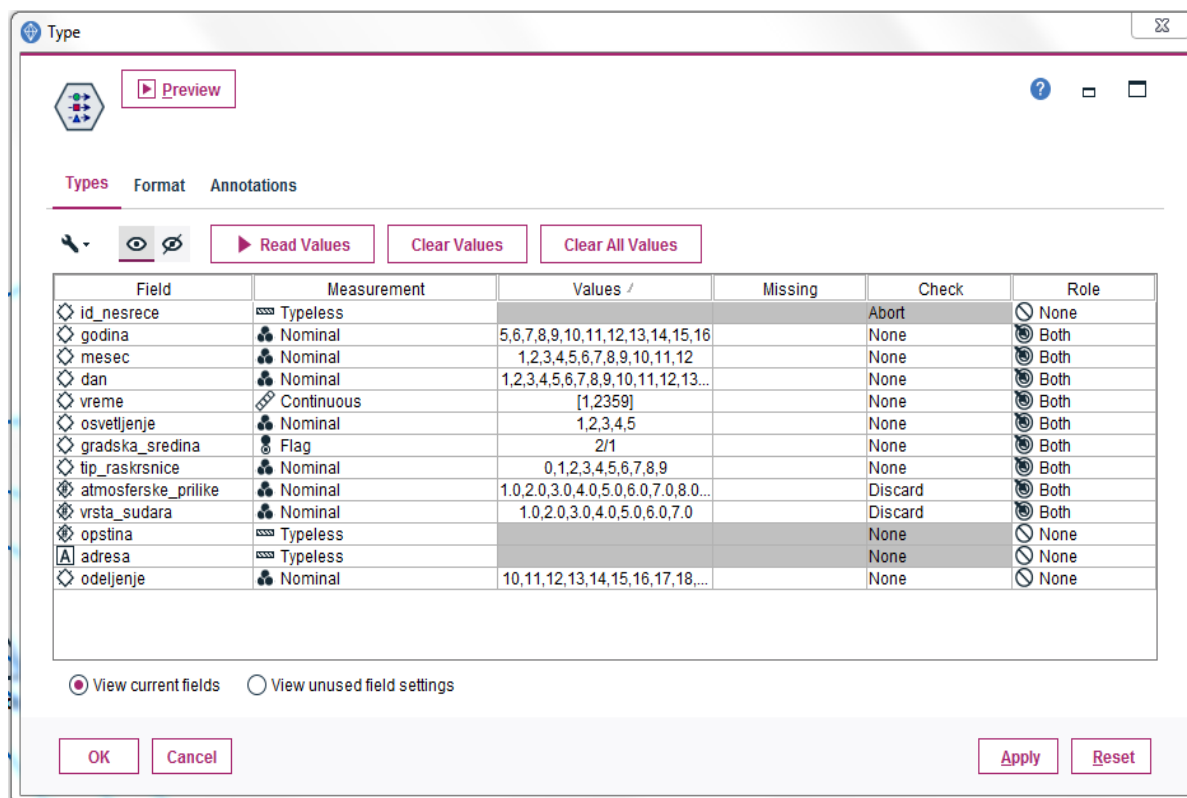


Figure 3: Sadržaj *Type* čvora za tabelu **characteristics**

vrednosti popunimo, odbacili smo i ove kolone. Kolona **env1** predstavlja predstavlja meru blizine školi, ali je zbog nejasnog kodiranja i ova kolona odbačena. Iako kolone **voie**, **vosp**, **lartpc**, **infra** i **nbv** nemaju puno nedostajućih vrednosti, gotovo svi slogovi uzimaju mali skup vrednosti za pomenute attribute pa

smo se odlučili da ni ove attribute ne koristimo u daljoj analizi. Za kolone **situ**, **prof**, **surf** i **plan** ćemo odbaciti slogove sa vrednošću nula za ove attribute. U koloni **larroust** se javljaju negativne vrednosti za širinu puta, pa ćemo i njih ukloniti.

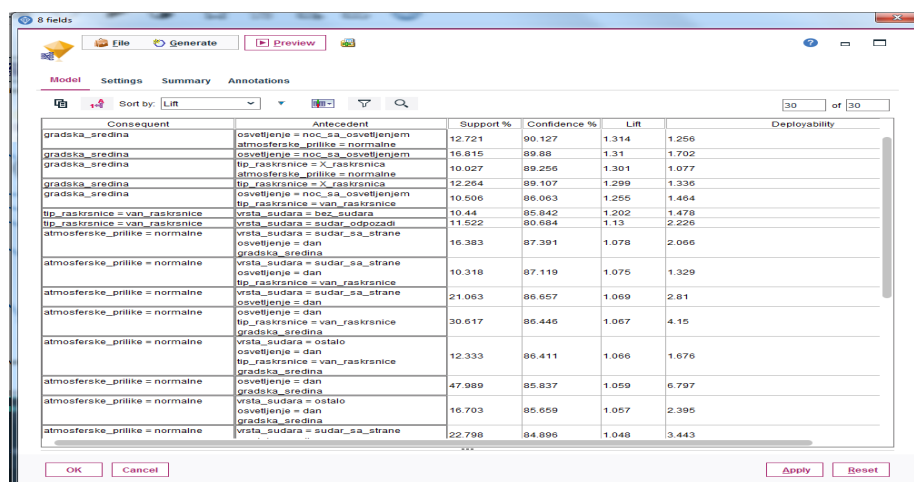
Skup podataka **holidays** smo odlučili da ne koristimo za dalju analizu, jer ne sadrži preterano korisne informacije.

4 Pravila pridruživanja

Nakon što smo obradili skupove podataka, hteli smo da na neke od relevantnih skupova primenimo algoritme *Apriori* i *Carma*, u nadi da ćemo uočiti neka zanimljiva pravila. Nakon primene pomenutih algoritama, cilj nam je bio da primenimo iste algoritme nad objedinjenim podacima.

4.1 Primena *Apriori* i *Carma* algoritama nad skupom characteristics

Iskoristili smo niz čvorova *Reclassify* kako bismo lakše tumačili kategoričke vrednosti. Iz tog skupa smo filtrirali one slogove čija je vrednost atributa *tip_raskrsnice* NA. Potom smo primenili *Apriori* sa podrazumevanim podešavanjima (minimalna podrška uzročnika je 10%, a minimalna pouzdanost pravila je 80%) i rezultati izvršavanja tog algoritma se mogu videti na slici. Posledice pravila se odnose na atmosferske prilike, tip raskrsnice i indikator da li se nesreća desila u gradu ili ne. Algoritam je uspeo da nadje 30 pravila, koja sve u svemu nisu zanimljiva. Naime, lift mera se kreće u opsegu od 0.998 do 1.314, što nam govori da su uzročnici u blagoj korelaciji sa posledicama. Kako bismo pripremili skup podataka za algoritam *Carma*, koristili smo čvor *SetToFlag*. *Carma* algoritam smo primenili sa istim parametrima kao i *Apriori*. Dobili smo 22 pravila, koja su gotovo identična onima koje smo dobili korišćenjem *Apriori* algoritma.



Consequent	Antecedent	Support %	Confidence %	Lift	Deplorability
gradska_sredina	osvetljenje = noc_sa_osvetljenjem	12.721	90.127	1.314	1.256
gradska_sredina	atmosferske_prilike = normalne	16.815	89.88	1.31	1.702
gradska_sredina	osvetljenje = noc_sa_osvetljenjem	10.027	89.256	1.301	1.077
gradska_sredina	tip_raskrsnice = X_raskrsnica	12.264	89.107	1.299	1.336
gradska_sredina	atmosferske_prilike = normalne	10.506	86.063	1.255	1.464
tip_raskrsnice = van_raskrsnice	osvetljenje = noc_sa_osvetljenjem	10.44	85.842	1.202	1.478
tip_raskrsnice = van_raskrsnice	vrata_sudara = sudar_sudara	11.522	80.684	1.13	2.226
atmosferske_prilike = normalne	vrata_sudara = sudar_sa_strane	16.383	87.391	1.078	2.066
atmosferske_prilike = normalne	osvetljenje = dan	10.318	87.119	1.075	1.329
atmosferske_prilike = normalne	vrata_sudara = sudar_sa_strane	21.063	86.657	1.069	2.81
atmosferske_prilike = normalne	tip_raskrsnice = van_raskrsnice	30.617	86.445	1.067	4.15
atmosferske_prilike = normalne	osvetljenje = dan	12.333	86.411	1.066	1.676
atmosferske_prilike = normalne	vrata_sudara = sudar	47.989	85.837	1.059	6.797
atmosferske_prilike = normalne	osvetljenje = dan	16.703	85.659	1.057	2.395
atmosferske_prilike = normalne	vrata_sudara = sudar_sa_strane	22.798	84.896	1.048	3.443

Figure 4: Apriori

Consequent	Antecedent	Support %	Confidence %	Lift	Deployability
gradska_sredina	atmosferske_prilike_normalne	12.721	90.127	1.314	1.256
gradska_sredina	osvetljenje_noc_sa_osvetljenjem	16.815	89.88	1.31	1.702
gradska_sredina	tip_raskrsnice_X_raskrsnica	12.264	89.107	1.299	1.336
atmosferske_prilike_normalne	gradska_sredina	16.383	87.391	1.078	2.066
atmosferske_prilike_normalne	vrsta_sudara_sudar_sa_strane	21.063	86.657	1.069	2.81
atmosferske_prilike_normalne	osvetljenje_dan	30.617	86.446	1.067	4.15
atmosferske_prilike_normalne	gradska_sredina	12.333	86.411	1.066	1.676
atmosferske_prilike_normalne	tip_raskrsnice_van_raskrsnice	47.989	85.837	1.059	6.797
atmosferske_prilike_normalne	osvetljenje_dan	16.703	85.659	1.057	2.395
atmosferske_prilike_normalne	gradska_sredina	22.798	84.896	1.048	3.443
atmosferske_prilike_normalne	vrsta_sudara_sudar_sa_strane	13.685	84.766	1.046	2.085
atmosferske_prilike_normalne	tip_raskrsnice_van_raskrsnice	68.678	84.346	1.041	10.751
atmosferske_prilike_normalne	osvetljenje_dan	48.801	84.21	1.039	7.706

Figure 5: Carma

Kao što se iz rezultata *Data Audit* čvora može primetiti, određene vrednosti nekih atributa dominiraju nad ostalim vrednostima, te stoga ne čudi što otkrivena pravila sadrže te vrednosti.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
godina		Nominal	5	16	—	—	—	12	839921
mesec		Nominal	1	12	—	—	—	12	839921
dan		Nominal	1	31	—	—	—	31	839921
vreme		Continuous	1	2359	1381.087	540.758	-0.488	—	839921
osvetljenje		Nominal	—	—	—	—	—	5	839921
gradska_sredina		Flag	1	2	—	—	—	2	839921
tip_raskrsnice		Nominal	—	—	—	—	—	10	839921
atmosferske_prilike		Nominal	—	—	—	—	—	9	839921
vrsta_sudara		Nominal	—	—	—	—	—	7	839921

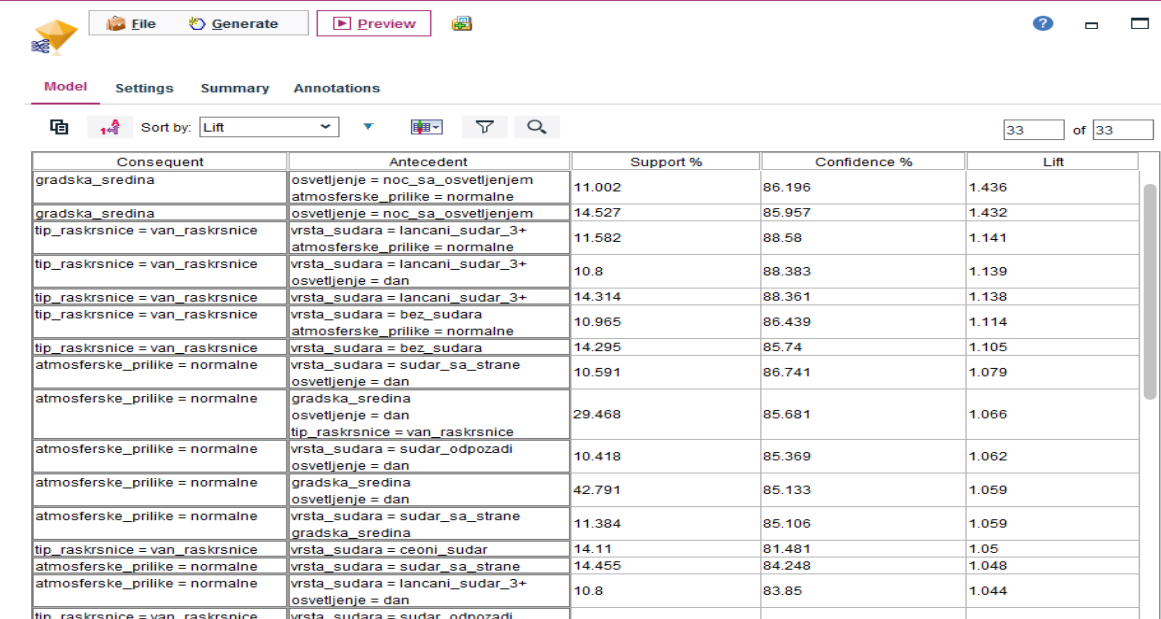
Figure 6: Možemo videti da se u kolonama *osvetljenje*, *tip_raskrsnice* i *atmosferske_prilike* u najvećem broju slučajeva javlja samo jedna vrednost, te ćemo te vrednosti pokušati da izbalansiramo. Još jedna kolona na koju ćemo primeniti balansiranje je *vrsta_sudara*.

U želji da izbor pravila bude pravedniji, odlučili smo da izbalansiramo skup podataka, tako što ćemo korišćenjem čvorova *Balance* na pojedinačne kolone ublažiti efekat dominantnih vrednosti (pomenute čvorove smo generisali uz pomoć distribucija odgovarajućih kolona). Nakon toga smo redom primenjivali *Apriori* algoritam za svaku izmenjenu kolonu. Potom smo eksperimentisali sa primenom *Apriori* algoritma na ulančane *Balance* čvorove. Neki od rezultata su predstavl-

jeni na narednim slikama.

tip_raskrsnice = van_raskrsnice	osvetljenje = noc_bez_osvetljenja	20.042	91.495	1.221	1.704
tip_raskrsnice = van_raskrsnice	osvetljenje = noc_bez_osvetljenja atmosferske_prilike = normalne	14.129	91.774	1.225	1.162
gradska_sredina	osvetljenje = noc_sa_osvetljenjem tip_raskrsnice = van_raskrsnice	12.364	85.211	1.402	1.829
gradska_sredina	osvetljenje = noc_sa_osvetljenjem	20.059	89.59	1.474	2.088
gradska_sredina	tip_raskrsnice = X_raskrsnica	10.605	89.611	1.475	1.102
gradska_sredina	osvetljenje = noc_sa_osvetljenjem atmosferske_prilike = normalne	15.254	89.79	1.478	1.557

Figure 7: Rezultat primene apriori algoritma na balansiranu kolonu *osvetljenje*. Izdvojena pravila jesu logična, ali nam ne otkrivaju puno interesantnih zaključaka.



Consequent	Antecedent	Support %	Confidence %	Lift
gradska_sredina	osvetljenje = noc_sa_osvetljenjem atmosferske_prilike = normalne	11.002	86.196	1.436
gradska_sredina	osvetljenje = noc_sa_osvetljenjem	14.527	85.957	1.432
tip_raskrsnice = van_raskrsnice	vrsta_sudara = lancani_sudar_3+ atmosferske_prilike = normalne	11.582	88.58	1.141
tip_raskrsnice = van_raskrsnice	vrsta_sudara = lancani_sudar_3+ osvetljenje = dan	10.8	88.383	1.139
tip_raskrsnice = van_raskrsnice	vrsta_sudara = lancani_sudar_3+	14.314	88.361	1.138
tip_raskrsnice = van_raskrsnice	vrsta_sudara = bez_sudara atmosferske_prilike = normalne	10.965	86.439	1.114
tip_raskrsnice = van_raskrsnice	vrsta_sudara = bez_sudara	14.295	85.74	1.105
atmosferske_prilike = normalne	vrsta_sudara = sudar_sa_strane osvetljenje = dan	10.591	86.741	1.079
atmosferske_prilike = normalne	gradska_sredina osvetljenje = dan tip_raskrsnice = van_raskrsnice	29.468	85.681	1.066
atmosferske_prilike = normalne	vrsta_sudara = sudar_odpozadi osvetljenje = dan	10.418	85.369	1.062
atmosferske_prilike = normalne	gradska_sredina osvetljenje = dan	42.791	85.133	1.059
atmosferske_prilike = normalne	vrsta_sudara = sudar_sa_strane gradska_sredina	11.384	85.106	1.059
tip_raskrsnice = van_raskrsnice	vrsta_sudara = ceoni_sudar	14.11	81.481	1.05
atmosferske_prilike = normalne	vrsta_sudara = sudar_sa_strane	14.455	84.248	1.048
atmosferske_prilike = normalne	vrsta_sudara = lancani_sudar_3+ osvetljenje = dan	10.8	83.85	1.044
tip_raskrsnice = van_raskrsnice	vrsta_sudara = sudar_odpozadi	11.807	80.966	1.043

Figure 8: Rezultat primene apriori algoritma na balansiranu kolonu *vrsta_sudara*. U ovom slučaju je pronađeno 33 pravila od kojih su ona sa najvećom lift merom logična ali i dalje nam ne daju upotrebljiviji uvid u zavisnosti među atributima.

Zaključujemo da balansiranje pojedinačnih kolona ne dovodi do željenih rezultata te smo probali sa ulančanim balansiranjem.

Model Settings Summary Annotations				
<div> <div>Sort by: Confidence %</div> <div>1 of 1</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
gradska_sredina	osvetljenje = noc_sa_osvetljenjem	20.299	85.784	1.505

Figure 9: Rezultat primene apriori algoritma na redom izbalansirane sve kolone skupa. Izdvojeno pravilo prema lift meri jeste zanimljivo ali je opet očekivano da u gradskoj sredini postoji osvetljenje koje je uključeno.

Model Settings Summary Annotations				
<div> <div>Sort by: Lift</div> <div>4 of 4</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
osvetljenje = dan	atmosferske_prilike = bljestavo_vreme	14.316	92.862	1.582
gradska_sredina	osvetljenje = noc_sa_osvetljenjem	19.314	87.437	1.319
gradska_sredina	vrsta_sudara = ostalo	17.067	82.009	1.237
gradska_sredina	atmosferske_prilike = lagana_kisa	11.662	81.935	1.236

Figure 10: Rezultat primene apriori algoritma na balansirane kolone *tip_raskrsnice*, *vrsta_sudara* i *atmosferske_prilike*.

Kako pokušaj sa balansiranjem nije prošao slavno, odlučili smo se da potpuno eliminišemo slogove koji imaju najzastupljeniju vrednost određenog atributa, i da na njega primenimo iste algoritme. Na ovaj način smo otkrili više pravila nego u prethodnim pokušajima, sa lift merama u opsegu od 0.673 do 1.421, ali sa malom pouzdanošću i podrškom. Rezultati se mogu videti na slici.

Model Settings Summary Annotations				
<div> <div>Sort by: Confidence %</div> <div>100 of 100</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
vrsta_sudara = bez_sudara	osvetljenje = noc_bez_osvetljenja atmosferske_prilike = normalne tip_raskrsnice = van_raskrsnice	6.993	38.917	1.421
vrsta_sudara = bez_sudara	tip_raskrsnice = kruzni_tok	2.202	38.626	1.41
vrsta_sudara = bez_sudara	osvetljenje = noc_bez_osvetljenja atmosferske_prilike = normalne	7.424	38.573	1.408
vrsta_sudara = bez_sudara	osvetljenje = noc_bez_osvetljenja tip_raskrsnice = van_raskrsnice	9.814	38.531	1.407
vrsta_sudara = bez_sudara	osvetljenje = noc_bez_osvetljenja	10.426	38.177	1.394
vrsta_sudara = sudar_odpozadi	tip_raskrsnice = T_raskrsnica osvetljenje = dan atmosferske_prilike = normalne	3.121	37.868	1.252
vrsta_sudara = sudar_odpozadi	tip_raskrsnice = T_raskrsnica osvetljenje = dan	3.745	37.572	1.243
vrsta_sudara = sudar_odpozadi	tip_raskrsnice = kruzni_tok	2.202	36.952	1.222
vrsta_sudara = sudar_odpozadi	atmosferske_prilike = normalne	4.209	36.385	1.203
vrsta_sudara = sudar_odpozadi	tip_raskrsnice = T_raskrsnica	5.236	35.776	1.183
vrsta_sudara = bez_sudara	atmosferske_prilike = lagana_kisa osvetljenje = noc_sa_osvetljenjem	2.162	33.791	1.233
vrsta_sudara = ceoni_sudar	tip_raskrsnice = X_raskrsnica atmosferske_prilike = normalne	4.698	33.639	1.29
vrsta_sudara = bez_sudara	osvetljenje = noc_sa_osvetljenjem tip_raskrsnice = van_raskrsnice	10.234	33.585	1.226

Figure 11: Nakon izbacivanja slogova koji bi ‘prigušili’ ostatak skupa, dobijeni su ovakvi rezultati. I dalje smatramo da ne postoje izuzetno zanimljiva pravila.

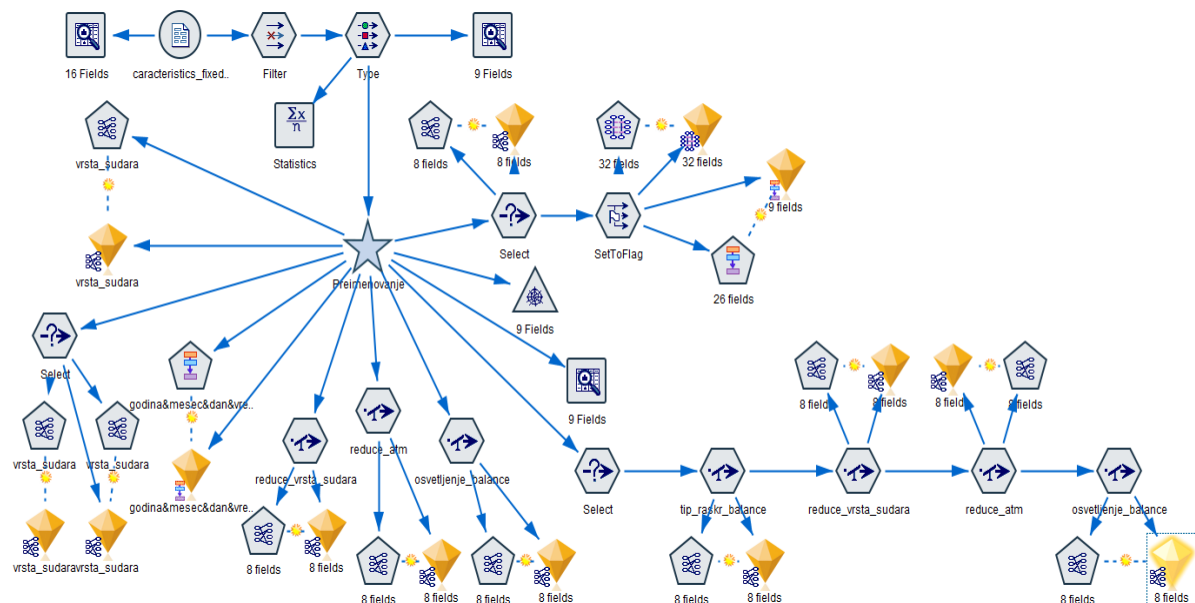


Figure 12: Prikaz celokupnog streama obrade skupa *characteristics.csv*

4.2 Primena *Apriori* i *Carma* algoritama nad skupom places

Nakon pomenutih koraka u preprocesiranju podataka ovog skupa, pokušali smo da primenimo apriori i carma algoritme kako bismo otkrili neke zanimljivosti, ali rezultati nisu bili preterano obećavajući. Naime, samo je carma čvor izdvojio pravila, ali ni ona nam nisu koristila za rezonovanje o skupu.

Types

Format

Annotations

Read Values

Clear Values

Clear All Values

Field	Measurement	Values	Missing	Check	Role
Num_Acc	Continuous	[20500000001,20...		None	None
kategorija_puta	Nominal	1,2,3,4,5,6,9		None	Both
tip_saobracaja	Nominal	0,1,2,3,4		None	Both
broj_traka	Ordinal	0,1,2,3,4,5,6,7,8,9,...		None	Both
nagib_puta	Nominal	0,1,2,3,4		None	Both
izgled_na_mapi	Nominal	0,1,2,3,4		None	Both
sirina_puta	Continuous	[-81,999]		None	Both
stanje_terena	Nominal	0,1,2,3,4,5,6,7,8,9		None	Both
pozicija_nesrece	Nominal	0,1,2,3,4,5		None	Both

Figure 13: Type čvor skupa *places*

Model Settings Summary Annotations				
<div> Sort by: Lift 281 of 281 </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
tip_saobracaja_2	kategorija_puta_3 broj_traka_2	23.433	89.944	1.41
broj_traka_2	kategorija_puta_3 tip_saobracaja_2	24.08	87.525	1.394
tip_saobracaja_2	kategorija_puta_3	29.032	82.942	1.3
broj_traka_2	kategorija_puta_3	29.032	80.712	1.285
tip_saobracaja_2	kategorija_puta_3 pozicija_nesrece_1	24.848	81.747	1.281
tip_saobracaja_2	kategorija_puta_4 broj_traka_2 izgled_na_mapi_1	27.632	81.423	1.276
tip_saobracaja_2	kategorija_puta_4 broj_traka_2 izgled_na_mapi_1 pozicija_nesrece_1	26.464	81.35	1.275
tip_saobracaja_2	kategorija_puta_4 broj_traka_2 stanje_terena_1	25.807	81.116	1.272
tip_saobracaja_2	kategorija_puta_4 broj_traka_2	31.347	81.043	1.27
tip_saobracaja_2	broj_traka_2 izgled_na_mapi_1 stanje_terena_1 pozicija_nesrece_1	38.613	80.901	1.268
tip_saobracaja_2	broj_traka_2 izgled_na_mapi_1 stanje_terena_1	41.628	80.845	1.267

Figure 14: Rezultat algoritma carma za skup *places*

4.3 Primena *Apriori* i *Carma* algoritama nad objedinjenim podacima

Uz pomoć čvora *Merge* smo spojili skupove *users*, *vehicles* i *characteristics* izbacujući redove sa nedostajućim vrednostima usput. Nakon toga smo izbacili kolone ... Izvršili smo dodatno uklanjanje besmislenih slogova i tako pripremljene podatke propustili kroz *Apriori* i *Carma* čvorove. U prvoj iteraciji smo koristili podrazumevane parametre za oba čvora. Rezultati su u nastavku.

Model Settings Summary Annotations				
Sort by: Lift				
Consequent	Antecedent	Support %	Confidenc...	Lift
zastita = 11	manevar = 1 kategorija_vozila = 7 vrsta_raskrsnice = 1 kategorija_ucesnika = 1 pokretna_prepreka = 2	10.287	90.531	1.564
zastita = 11	manevar = 1 kategorija_vozila = 7 vrsta_raskrsnice = 1 osvetljenje = 1 kategorija_ucesnika = 1	10.007	89.416	1.545
zastita = 11	manevar = 1 kategorija_vozila = 7 vrsta_raskrsnice = 1 kategorija_ucesnika = 1	14.147	89.119	1.539
zastita = 11	manevar = 1 kategorija_vozila = 7 osvetljenje = 1 kategorija_ucesnika = 1 pokretna_prepreka = 2	11.629	88.83	1.534
zastita = 11	manevar = 1 kategorija_vozila = 7 vrsta_raskrsnice = 1 kategorija_ucesnika = 1 atmosferske_prilike = 1.0	11.147	88.799	1.534
zastita = 11	pol kategorija_vozila = 7 vrsta_raskrsnice = 1 kategorija_ucesnika = 1	11.426	88.695	1.532

Figure 15: Rezultati apriori algoritma sa podrazumevanim parametrima nad objedinjenim skupom.

Model Settings Summary Annotations				
Sort by: Lift				
Consequent	Antecedent	Support %	Confidence %	Lift
vrsta_raskrsnice_1 kategorija_ucesnika_1 kategorija_vozila_7	zastita_11 osvetljenje_1	41.425	49.011	1.518
ozbilnost_1 kategorija_vozila_7 pokretna_prepreka_2	atmosferske_prilike_1.0 zastita_11	46.355	43.773	1.515
atmosferske_prilike_1.0 zastita_11	ozbilnost_1 kategorija_vozila_7 pokretna_prepreka_2	28.901	70.207	1.515
kategorija_vozila_7 pokretna_prepreka_2 osvetljenje_1	atmosferske_prilike_1.0 kategorija_ucesnika_1 zastita_11	38.314	54.83	1.513
atmosferske_prilike_1.0 kategorija_ucesnika_1 zastita_11	kategorija_vozila_7 pokretna_prepreka_2 osvetljenje_1	36.238	57.972	1.513
zastita_11	vrsta_raskrsnice_1 kategorija_ucesnika_1 kategorija_vozila_7 osvetljenje_1	23.189	87.554	1.512
vrsta_raskrsnice_1 kategorija_ucesnika_1 kategorija_vozila_7 osvetljenje_1	zastita_11	57.893	35.07	1.512
atmosferske_prilike_1.0 kategorija_ucesnika_1 kategorija_vozila_7 osvetljenje_1	ozbilnost_1	45.988	43.699	1.508
ozbilnost_1	atmosferske_prilike_1.0			

Figure 16: Rezultati carma algoritma sa podrazumevanim parametrima nad objedinjenim skupom.

Primećujemo da su prisutna pravila sa solidnom lift merom od oko 1.5 i velikom podrškom za oba algoritma. Odmah se Primećuje i ogromna razlika u količini pronadjenih pravila. Većina pravila pronadjenih apriori algoritmom kao posledicu imaju nošenje zaštitnog pojasa u trenutku nesreće. Kod carma algoritma pravila su malo raznolikija ali i dalje se u pravilima javljaju većinski dominantne vrednosti odgovarajućih kolona. U nastavku ćemo pokušati da ispitamo vezu izmedju nekih atributa za koje mislimo da mogu proizvesti interesantna pravila.

Najpre smo hteli da ispitamo postoje li veze izmedju atmosferskih prilika, vrste sudara, kategorija učesnika, ozbiljnosti povreda i korišćene zaštitne opreme. Kroz više iteracija smo se zaustavili na minimalnoj podršci od 5.0% i minimalnoj pouzdanosti od 60.0%. Pravila dobijena apriorijem kojima je lift mera najveća uglavnom za posledicu imaju nošenje zaštitnog pojasa, pa nam ovo nije preterano zanimljivo.

Model Settings Summary Annotations				
<div> </div> <div> Sort by: Lift </div> <div> 70 of 70 </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita = 11	vrsta_sudara = 4.0 kategorija_ucesnika = 1	5.107	82.984	1.433
zastita = 11	vrsta_sudara = 4.0	6.269	81.95	1.416
zastita = 11	vrsta_sudara = 4.0 atmosferske_prilike = 1.0	5.054	81.747	1.412
zastita = 11	vrsta_sudara = 2.0 ozbiljnost = 1 kategorija_ucesnika = 1	5.504	81.219	1.403
zastita = 11	vrsta_sudara = 2.0 ozbiljnost = 1	6.127	80.487	1.39
zastita = 11	vrsta_sudara = 2.0 ozbiljnost = 1 atmosferske_prilike = 1.0	5.058	80.415	1.389
zastita = 11	vrsta_sudara = 3.0 ozbiljnost = 1 osvetljenje = 1 kategorija_ucesnika = 1	10.87	80.384	1.388
zastita = 11	vrsta_sudara = 3.0 ozbiljnost = 1 osvetljenje = 1 kategorija_ucesnika = 1 atmosferske_prilike = 1.0	9.431	80.352	1.388
zastita = 11	vrsta_sudara = 3.0 ozbiljnost = 1 kategorija_ucesnika = 1	14.832	79.956	1.381
zastita = 11	vrsta_sudara = 3.0 ozbiljnost = 1 kategorija_ucesnika = 1	12.493	79.869	1.38

Figure 17: Rezultati apriori algoritma primenjenog na atmosferske prilike, vrstu sudara, kategoriju učesnika, ozbiljnost povreda i korišćenu zaštitnu opremu.

Carma algoritam je dao malo zanimljivije rezultate. Naime, pronadjena su tri pravila koja lift merom odstupaju od jedinice. Pravilo *žensko + korišćen je zaštitni pojas* ima lift meru od 1.128 što nam govori da postoji blaga pozitivna korelacija izmedju ženskog pola i vezivanja zaštitnog pojasa. S druge strane, na osnovu lift mere manje od 1, dolazimo do zaključka da ukoliko je došlo do blage ozlede, možemo spekulirati da se ne radi o vozaču. Slično tome, ukoliko je učesnik nezgode žensko, manja je šansa da je vozač nego neka druga kategorija učesnika. Za ovaj čvor smo koristili minimalnu podršku od 20.0% i minimalnu pouzdanost od 60.0%.

Model Settings Summary Annotations				
<div> <div>Sort by: Lift</div> <div>6 of 6</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita_11	pol	33.073	65.329	1.128
kategorija_ucesnika_1	vrsta_sudara_3.0	37.228	83.841	1.086
zastita_11	kategorija_ucesnika_1	77.237	61.654	1.065
kategorija_ucesnika_1	zastita_11	57.893	82.255	1.065
kategorija_ucesnika_1	ozbiljnost_4	34.331	67.379	0.872
kategorija_ucesnika_1	pol	33.073	62.211	0.805

Figure 18: Rezultati carma algoritma primenjenog na atmosferske prilike, vrstu sudara, kategoriju učesnika, ozbiljnost povreda i korišćenu zaštitnu opremu.

Sledeći pokušaj se odnosio na uticaj zaštitne opreme i vrste sudara na ozbiljnost povreda. Carma čvor je ostavljen sa podrazumevanim vrednostima (po 20.0% za oba parametra). Čvoru apriori smo iterativno smanjivali parametre, ali ni za minimalnu podršku od 5.0% i minimalnu pouzdanost od 5.0% nije uspeo da pronadje nijedno pravilo.

Model Settings Summary Annotations				
<div> <div>Sort by: Lift</div> <div>4 of 4</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita_11	ozbiljnost_1	45.988	79.114	1.367
ozbiljnost_1	zastita_11	57.893	62.846	1.367
vrsta_sudara_3.0	zastita_11	57.893	34.989	0.94
zastita_11	vrsta_sudara_3.0	37.228	54.411	0.94

Figure 19: Rezultati carma algoritma primenjenog na vrstu sudara, zaštitu i ozbiljnost povrede. Jedini zaključak koji možemo izvesti je da korišćenje zaštitnog pojasa pozitivno korelira sa odsustvom ozlede. Podrška i lift mera ostala dva pravila nam govori da ta pravila nisu dovoljno interesantna.

Nakon toga smo se pitali da li kategorija učesnika ima veze sa korišćenom zaštitnom opremom.

Model Settings Summary Annotations				
<div> <div>Sort by: Lift</div> <div>3 of 3</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita = 21	kategorija_ucesnik...	83.396	24.682	1.078
zastita = 11	kategorija_ucesnik...	16.579	74.705	1.055
zastita = 11	kategorija_ucesnik...	83.396	70.047	0.989

Figure 20: Rezultati apriori algoritma primenjenog na kategoriju učesnika i korišćenu zaštitnu opremu. Iz sličnog razloga kao i za apriori odbacujemo ova pravila kao nekorisna.

Pitali smo se da li godine učesnika imaju veze sa korišćenjem zaštitne opreme. Najpre smo godine, koje su se nalazile u intervalu 1896 do 2016, ograničili na interval od 1920 do 2006 i podelili ih u 4 kategorije korišćenjem *binning* čvora.

Model Settings Summary Annotations				
<div> Sort by: Lift 4 of 4 </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita_21	kategorija_ucesnik...	83.396	24.682	1.078
kategorija_ucesnik...	zastita_21	22.891	89.918	1.078
kategorija_ucesnik...	zastita_11	70.801	82.507	0.989
zastita_11	kategorija_ucesnik...	83.396	70.047	0.989

Figure 21: Rezultati carma algoritma primenjenog na kategoriju učesnika i korišćenu zaštitnu opremu. Lift mera je veoma bliska jedinici te ne smatramo pravila interesantnim.

Nakon nekoliko pokretanja apriori čvora, došli smo do sledećih podešavanja: minimalna podrška od 5.0%, minimalna pouzdanost 20.0%.

Model Settings Summary Annotations				
<div> Sort by: Lift 5 of 5 </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita = 21	godina_rođenja_BIN = godine_manje_od_21	28.857	27.427	1.461
zastita = 11	godina_rođenja_BIN = godine_od_42_do_65	21.862	65.776	1.131
zastita = 11	godina_rođenja_BIN = godine_vece_od_65	6.79	63.902	1.099
zastita = 11	godina_rođenja_BIN = godine_od_21_do_42	42.49	61.276	1.054
zastita = 11	godina_rođenja_BIN = godine_manje_od_21	28.857	46.459	0.799

Figure 22: Rezultati apriori algoritma primenjenog na kategoriju starosti učesnika i korišćenu zaštitnu opremu. Dolazimo do zaključka da učesnici mlađji od 21 godine koriste kacigu kao i da ne nose pojas (doduše oba ova pravila imaju malu pouzdanost).

Algoritam carma sa minimalnom podrškom od 20.0% i minimalnom pouzdanošću 20.0% nije pronašao zanimljiva pravila (lift mera je veoma bliska jedinici).

Poslednji pokušaj ticao se traženja povezanosti izmedju ozbiljnosti povrede, izgleda puta na mapi, pozicije nesreće i zaštitne opreme. Apriori algoritam je korišćen sa parametrima 5.0% i 40.0% za minimalnu podršku i minimalnu pouzdanost, redom. Carma čvor smo koristili sa podrazumevanim parametrima.

Model Settings Summary Annotations				
<div> <div> <div></div> <div>1</div> </div> <div>Sort by: Lift</div> <div> <div></div> <div></div> <div></div> </div> <div>15 of 15</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita = 11	ozbiljnost = 1 pozicija_nesrece = 1	43.36	89.995	1.435
zastita = 11	ozbiljnost = 1 izgled_na_mapi = 1 pozicija_nesrece = 1	38.251	89.764	1.431
zastita = 11	ozbiljnost = 1	44.293	89.757	1.431
zastita = 11	ozbiljnost = 1 izgled_na_mapi = 1	39.052	89.528	1.428
zastita = 11	izgled_na_mapi = 3 pozicija_nesrece = 1	5.914	71.269	1.136
zastita = 11	izgled_na_mapi = 3	6.058	70.908	1.131
zastita = 11	izgled_na_mapi = 2 pozicija_nesrece = 1	5.647	68.909	1.099
zastita = 11	izgled_na_mapi = 2	5.815	68.591	1.094
zastita = 11	pozicija_nesrece = 1	97.863	62.962	1.004
zastita = 11	izgled_na_mapi = 1 pozicija_nesrece = 1	85.299	61.893	0.987
zastita = 11	izgled_na_mapi = 1	87.101	61.649	0.983
zastita = 11	ozbiljnost = 4 pozicija_nesrece = 1	35.724	44.039	0.702
zastita = 11	ozbiljnost = 4	36.449	43.718	0.697
zastita = 11	ozbiljnost = 4 izgled_na_mapi = 1 pozicija_nesrece = 1	31.757	42.051	0.671
zastita = 11	ozbiljnost = 4 izgled_na_mapi = 1	32.377	41.738	0.666

Figure 23: Rezultati apriori algoritma primenjenog na ozbiljnost povrede, izgled puta na mapi, poziciju nesreće i zaštitnu opremu. Najveći deo pravila kao posledicu ima korišćenje zaštitnog pojasa što je i očekivano s obzirom na broj slogova u kome se pojas javlja, ali smatramo da ovo pravilo ne pruža nikakav značajan zaključak.

Model Settings Summary Annotations				
<div> <div> <div></div> <div>1</div> </div> <div>Sort by: Lift</div> <div> <div></div> <div></div> <div></div> </div> <div>84 of 84</div> </div>				
Consequent	Antecedent	Support %	Confidence %	Lift
zastita_11 izgled_na_mapi_1	ozbiljnost_1 pozicija_nesrece_1	43.36	79.188	1.475
ozbiljnost_1 pozicija_nesrece_1	zastita_11 izgled_na_mapi_1	53.697	63.944	1.475
zastita_11 izgled_na_mapi_1	ozbiljnost_1	44.293	78.936	1.47
ozbiljnost_1	zastita_11 izgled_na_mapi_1	53.697	65.112	1.47
zastita_11 pozicija_nesrece_1 izgled_na_mapi_1 ozbiljnost_1	zastita_11	44.293	77.52	1.468
	pozicija_nesrece_1	52.794	65.037	1.468
	izgled_na_mapi_1			
ozbiljnost_1 pozicija_nesrece_1	zastita_11	62.71	62.226	1.435
zastita_11	ozbiljnost_1 pozicija_nesrece_1	43.36	89.995	1.435
zastita_11	pozicija_nesrece_1 izgled_na_mapi_1	38.251	89.764	1.431
ozbiljnost_1 pozicija_nesrece_1 izgled_na_mapi_1	zastita_11	62.71	54.753	1.431
zastita_11	ozbiljnost_1	44.293	89.757	1.431
ozbiljnost_1	zastita_11	62.71	63.396	1.431
zastita_11 pozicija_nesrece_1	ozbiljnost_1	44.293	88.1	1.43

Figure 24: Rezultati carma algoritma primenjenog na ozbiljnost povrede, izgled puta na mapi, poziciju nesreće i zaštitnu opremu. Dobijena pravila imaju dobru lift meru i logična su. Na primer, korišćenje pojasa na pravom putu povezana je sa odsustvom povreda.

5 Zaključak

Pravila pridruživanja predstavljaju moćan alat za uočavanje nekih krajnje neočekivanih zavisnosti medju podacima. Medjutim, iako su algoritmi za nalaženje pravila jednostavni za implementaciju, rezultati dosta zavise i od samih podataka. Naime, u poslednje vreme se javljaju radovi koji predlažu naprednije tehnike pretprocesiranja kao preduslov za izvlačenje boljih pravila. U našem slučaju, skup podataka je bio dosta neuravnotežen i smatramo to jednim od važnijih faktora za donekle neuspelo pronalaženje interesantnijih pravila.