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 1, 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 atribute svake od njih.

- `caracteristics.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 - čeoni 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`

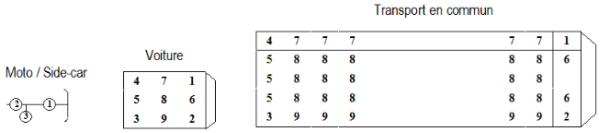
- **ds** : datum nesreće u formatu godina-mesec-dan
- **holiday** : naziv praznika

• `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 rodjenja - 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 - neodredjeno

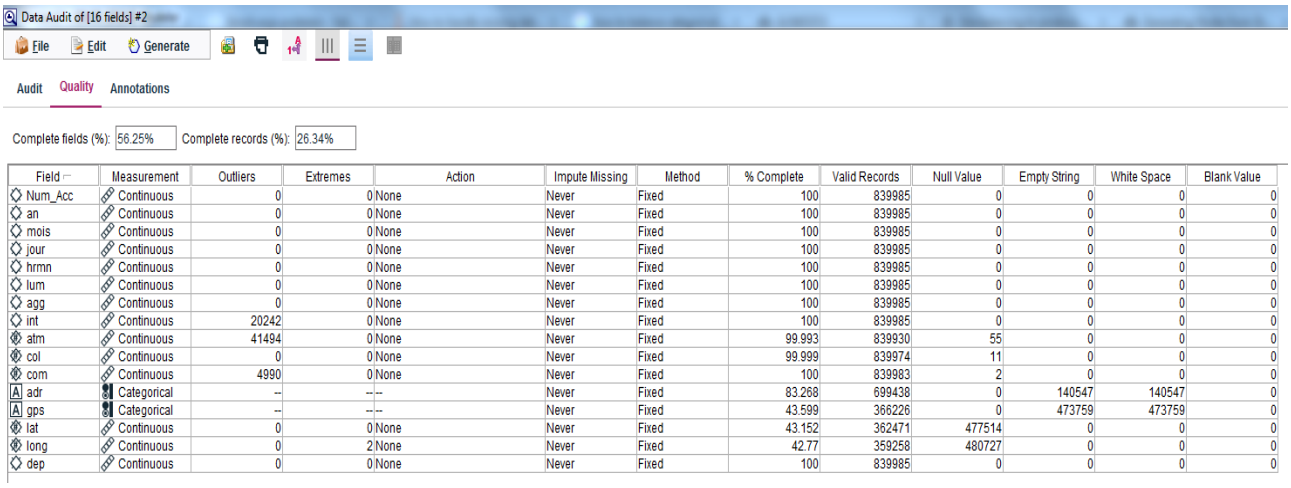
– **locp** : pozicija pešaka [1-8], kodirano na sledeći način:

    - 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 - neodredjeno
    - 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 odredjuje 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 smislena, 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`.


Data Audit of [16 fields] #2

File Edit Generate

Audit Quality Annotations

Complete fields (%): 56.25% Complete records (%): 26.34%

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
hrmn	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 neodredjenih 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 odredjenih polja sa zadatim opisom reprezentacije tog atributa, tako da smo te nekonzistentne vrednosti preimenovali u `NA` (neodredjenu 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 nedostajuće i neodredjene vrednosti možemo da odbacimo. Nismo naišli ni na kakve nepravilnosti koje iziskuju podrobnije procesiranje.

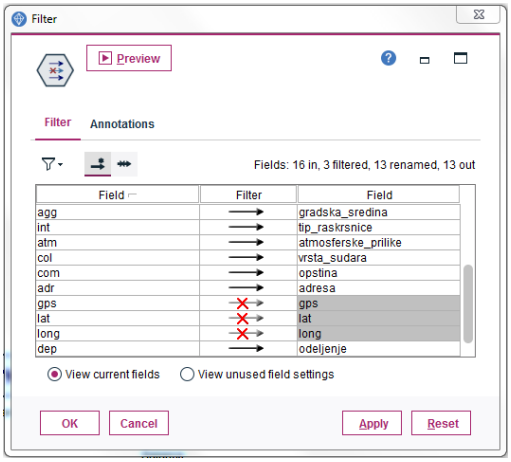


Figure 2: Sadržaj *Filter* čvora za tabelu `characteristics`

Skup podataka `places`

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 svaki 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.

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

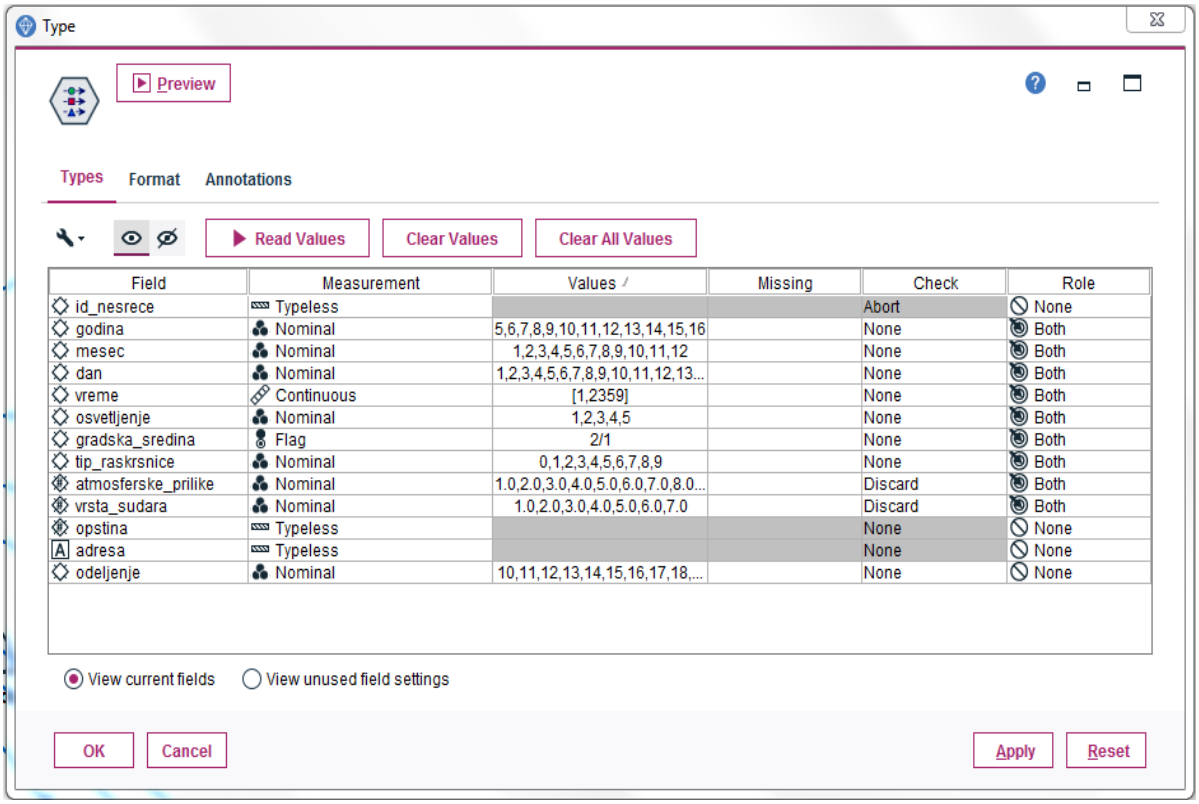


Figure 3: Sadržaj *Type* čvora za tabelu `characteristics`

toga smo redom primenjivali *Apriori* algoritam za svaku izmenjenu kolonu. Potom smo eksperimentisali sa primenom *Apriori* alogritma na ulančane *Balance* čvorove i na čvorove primenjene na pojedinačne kolone. Neki od rezultata su predstavljeni na narednim slikama.

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