Istraživanje podataka 1 - vežbe 12, 2020.

1 Pravila pridruživanja u SPSS Modeleru

1.1 Čvor Association Rules

Za izdvajanje pravila pridruživanja u IBM SPSS Modeleru može da se koristi čvor **Association Rules** koji se nalazi u paleti *Modeling* u delu *Association*. U odnosu na čvor **Apriori** čvor **Association Rules**:

- ne radi sa podacima u transakcionom obliku
- ima više opcija za postavljanje ograničenja koja se koriste pri izdvajanju zanimljivih pravila pridruživanja

Atributi u tabeli koja se koristi za određivanje pravila pridruživanja mogu biti različitih tipova. Jedan red je jedna transakcija. Svaka vrednost u kategoričkom atributu se posmatra kao jedna stavka. Nad numeričkim atributima se vrši diskretizacija i svaka nastala grupa se posmatra kao jedna stavka.

1.1.1 Parametri čvora Association Rules

- Odeljak *Fields*
  - *Use predefined roles* - koriste se uloge atributa koje su dodeljene pre korišćenja čvora **Apriori**. Prema dodeljenim ulogama, atributi će biti raspoređeni po listama *Both (Condition or Prediction)*, *Prediction only*, *Condition only*.
  - *Use custom field assignments* - definišu se uloge atributa. Korisnik dodeljuje uloge atributima razvrstavanjem po listama. Vrednosti atributa u listi *Both (Condition or Prediction)* mogu da se pojave u telu ili glavi pravila. Vrednosti atributa u listi *Prediction only* mogu da se pojave u glavi pravila, a vrednosti atributa u listi *Condition only* mogu da se pojave u telu pravila pridruživanja.
- Odeljak *Build options*
  - deo *Rule building* omogućava izbor algoritma i zadavanje ograničenja koja moraju da zadovolje pravila pridruživanja koja će biti u modelu
    - *Maximum conditions* - maksimalan broj stavki u telu jednog pravila
    - *Maximum predictions* - maksimalan broj stavki u glavi jednog pravila
    - *Algorithm* - algoriatm koji će se primeniti
    - *Only true values for flags* - ako su podaci o transakcijama u tabelarnom obliku sa binarnim atributima, uzima se u obzir samo pojavljivanje stavke u transakciji.

    * *Rule criterion* - postavljanje praga za različite mere kvaliteta koje moraju da zadovolje izdvojena pravila pridruživanja
    * *Exclude rules* - U nekim slučajevima je povezanost između dve ili više stavki poznata, i u takvim slučajevima je korisno isključiti pravila u kojima jedna stavka predviđa pojavljivaje druge. Pomoću ove opcije mogu se definisati poznate povezanosti, da se takva pravila ne bi izdvojila.
    * *Maximum number of rules* - maksimalan broj pravila pridruživanja koja će biti izdvojena
    * *Rule criterion for top N* - mera kvaliteta prema kojoj se određuje koja su pravila pridruživanja najbolja
    * *Only true values for flags* - uzeti u obzir samo vrednosti **tačno** u binarnim atributima. Ako ova opcija nije izabrana može se izdvojiti i pravilo oblika $A{=}T$ *&* $B{=}F \rightarrow C{=}T$.
  - deo *Transformations* omogućava definisanje parametara koji se koriste pri automatskoj transformaciji skupa pre izdvajanja pravila pridruživanja
    * *Binning* - broj grupa za diskretizaciju numeričkih atributa. Svi numerički atributi se automatski diskretizuju i pravi se zadati broj grupa jednake širine.
  - deo *Output* omogućava definisanje izgleda konačnog izveštaja o izdvojenim pravilima pridruživanja kada se napravi model
    * *Rules tables* - izbor mera kvaliteta. Za svaku izabranu meru kvaliteta biće prikazana posebna tabela sa pravilima pridruživanja koja su uređena prema vrednosti za tu meru. Pomoću opcije *Rules to display*, zadaje se broj pravila koja će biti izdvojena po tabeli.
    * *Model information tables* - izbor informacija koje će biti prikazane u izveštaju
      · *Field Transformations* - podaci o izvršenim transformacijama nad atributima (npr. o diskretizaciji numeričkih atributa)
      · *Records Summary* - podaci o broju slogova koji su korišćeni za pravljenje modela, kao i broju slogova koji su isključeni
      · *Rule Statistics* - prikaz osnovnih statistika izračunatih nad vrednostima osnovnih mera kvaliteta za izdvojena pravila pridruživanja. Za jednu meru kvlaiteta izdvaja se: najmanja, najveća, srednja vrednost i standardna devijacija.
      · *Most Frequent Values* - prikaz podataka o najčešćim stavkama u skupu
      · *Most Frequent Fields* - prikaz podataka o atributima čije se vrednosti najčešće javljaju kao stavke u pravilima pridruživanja

- odeljak *Model Options* se koristi za postavljanje uslova koji moraju da važe pri određivanju najboljih pravila pridruživanja koja veže u jednoj transakciji. Opcije su:
  - *Maximum number of predictions* - broj pravila koja će biti dodeljena svakom redu (transakciji) u tabeli sa skupom podataka. Pri dodeli pravila biraju se najbolja pravila prema izabranoj meri preko liste *Rule Criterion*.
  - *Allow repeat predictions* - da li je pri izboru najboljih pravila dozvoljeno ponavljanje glave ili ne. Ako jeste, onda je moguće imati više najboljih pravila sa istom glavom, a ako nije onda izdvojena najbolja pravila moraju imati različite glave.

- *Only score rules when predictions are not present in the input* - glava pravila ne sme da se pojavljuje u transakciji
- *Only score rules when predictions are present in the input* - glava pravila mora da se pojavljuje u transakciji
- *Score all rules* - glava pravila može, a i ne mora da se pojavljuje u transakciji

1.2 Model dobijen primenom čvora Association Rules

Model koji se dobija kao rezultat korišćenja čvora **Association Rules** je prikazan u obliku dijamanta u radnom toku. Podaci u modelu su prikazani u obliku tabela koje se mogu jednostavno sačuvati u različitim formatima, što je korisno pri pravljenju izveštaja. U modelu u odeljku *Model* se prikazuju tabele koje su izabrane preko opcija u čvoru **Association Rules**. Prikazane tabele u modelu su podeljene u tri dela:

- *Model Information* koji sadrži podatke o zadatim pragovima za mere kvaliteta, izvršenim transformacijama nad atributima i broju transkacija koje su korišćene za izdvajanje pravila pridruživanja.
- *Evaluation* sadrži statistike o izsdvojenim pravilima i najčešćim stavkama u transakcijama.
- *Interpretation* sadrži tabele sa izdvojenim pravilima pridruživanja uređenim prema vrednostima izabranih mera kvaliteta.

Odeljak *Model Options* se koristi za postavljanje uslova koji moraju da važe pri dodeli najboljih pravila pridruživanja za svaku transakciju. Opcije su iste kao u odeljku *Model Options* čvora **Association Rules**.

Klikom na dugme *Preview* za svaku transakciju (red) u skupu se prikazuju podaci o najboljim pravilima koja važe u toj transakciji prema zadatim uslovima. Za jedno pravilo se prikazuje: glava pravila, vrednost zadate mere za izbor najboljih pravila i ID pravila. Celo pravilo se preko identifikatora može pronaći u delu *Interpretation* u odeljku *Model*.

2 Zadaci IBM SPSS Modeleru

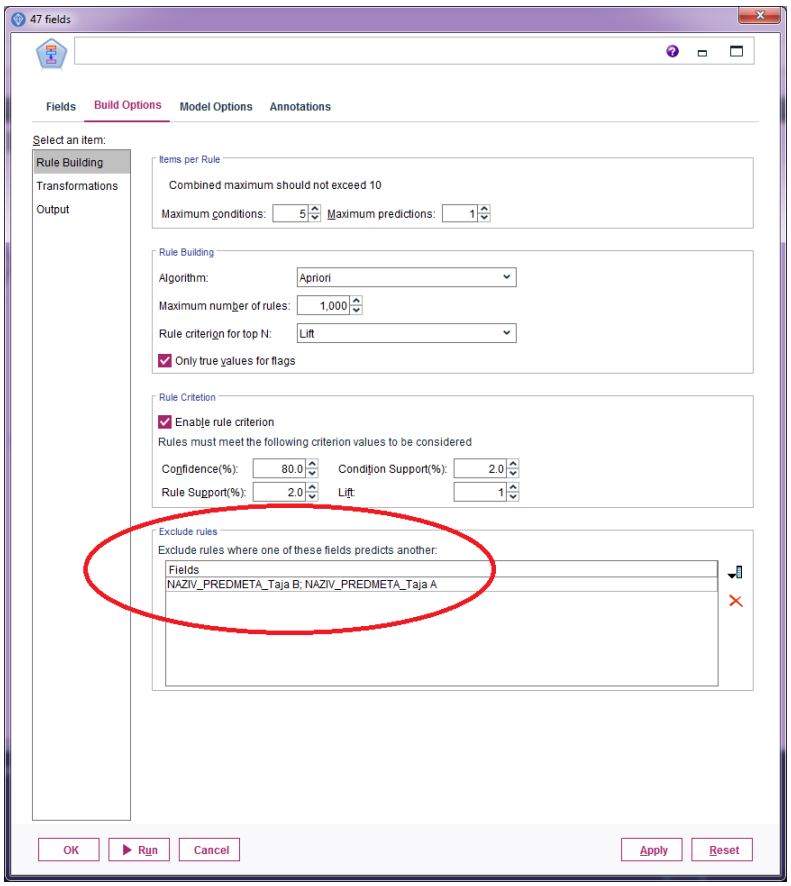
1. Primenom pravila pridruživanja proveriti da li postoji zavisnost među upisanim izbornim predmetima studenata smera Informatika i Računarstvo i informatika.
   Skup u datoteci *izborni_predmeti.xlsx* sadrži podatake o upisanim izbornim predmetima studenata. U jednom redu je indeks studenta, naziv jednog od njegovih izbornih predmeta koje je upisao i naziv smera koji studira.

   Rešenje

   Radni tok: **pravila_pridruzivanja_izborni_predmeti.str**

   U oviru vežbi 11 primenom čvora **Apriori** urađen je ovaj primer. Za detalje o pripremi podataka, pravljenju modela, zadatim parametrima i zaključcima pogledati rešenje primera 2, vežbe 11.

U modelu dobijenom primenom čvora **Apriori** postoji nekoliko pravila koja su zanimljiva prema merama kvaliteta za pravila pridruživanja, a koja sadrže predmete Taja A i Taja B. Pošto se zna da je predmet Taja A uslovni za predmet Taja B, ova pravila nisu zanimljiva. Ako se koristi čvor **Association Rules** za izdvajanje pravila pridruživanja, ovo znanje se može uključiti pri pravljenju modela pomoću opcije *Exclude rules* (slika 1), čime takva pravila neće biti izdvojena. U čvoru modela, u odeljku *Model*, izdvojena je tabela sa pravilima pridruživanja koja su uređena prema vrednosti za Lift meru. Izdvojena su ista pravila kao i kada se koristi čvor **Apriori**, samo bez pravila sa predmetima Taja A i Taja B.



Slika 1: Postavljanje parametara za u čvoru *Association Rules*

2. Korišćenjem čvora **Association Rules** izdvojiti pravila pridruživanja iz skupa u datoteci *smer_predmet_rok.csv* koji sadrži podatke o uspešnosti polaganja ispita studenata određenog smera u određenom roku. Atributi su:

   - *NAZIV_SMERA* - naziv smera studenata koji su polagali ispit
   - *NAZIV_PREDMETA* - naziv predmeta koji je polagan
   - *GODINA_ROKA* - godina u kojoj je održan ispit
   - *OZNAKA_ROKA* - oznaka ispitnog roka
   - *PROSEK* - prosečna ocena na položenim ispitima
   - *BROJ_POLAGANJA* - koliko puta su u proseku studenti koji su izašli na ispit polagali taj predmet

   U radnom toku postaviti da važe sledeći uslovi:

- koriste se svi atributi osim *GODINA_ROKA*
- numerički atributi se dele u 4 grupe
- izdvaja se najviše 500 najboljih pravila pridruživanja prema Lift meri
- minimalna podrška pravila je 5%
- minimalna pouzdanost pravila je 60%
- izdvajaju se pravila koja su zanimljiva prema Lift meri

Odgovoriti na pitanja:

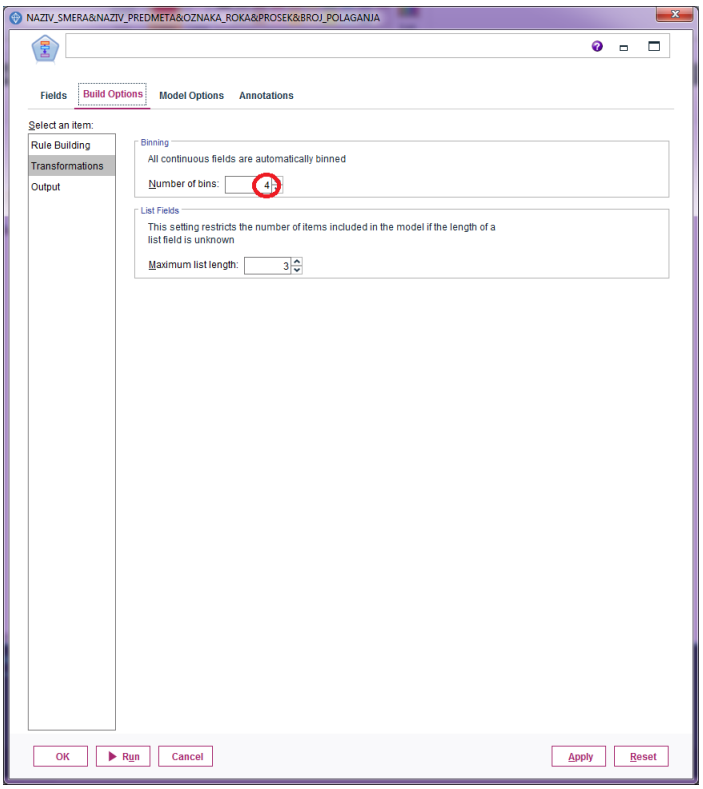
- Među izdvojenim pravilima pridruživanja, koje pravilo je najzanimljivije?
- Pronaći najbolje pravilo prema Lift meri za 8. transakciju u skupu kada se stavke koje su u glavi pravila javljaju u transakciji.
- Koja stavka je najčešća u skupu i kolika joj je podrška?

Rešenje

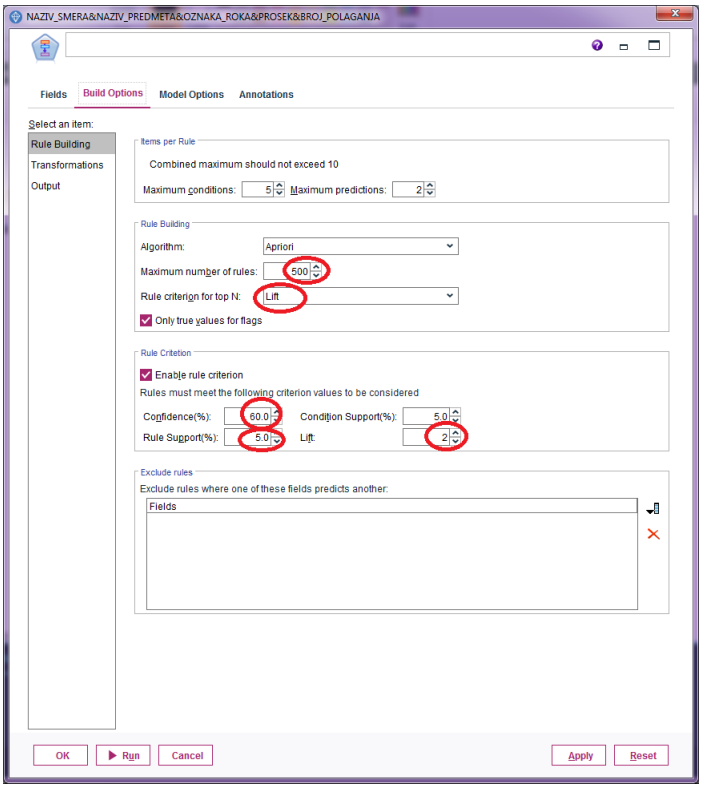
Radni tok: **pp_smer_predmet_rok.str**

Postavke opcija u čvorovima u radnom toku:

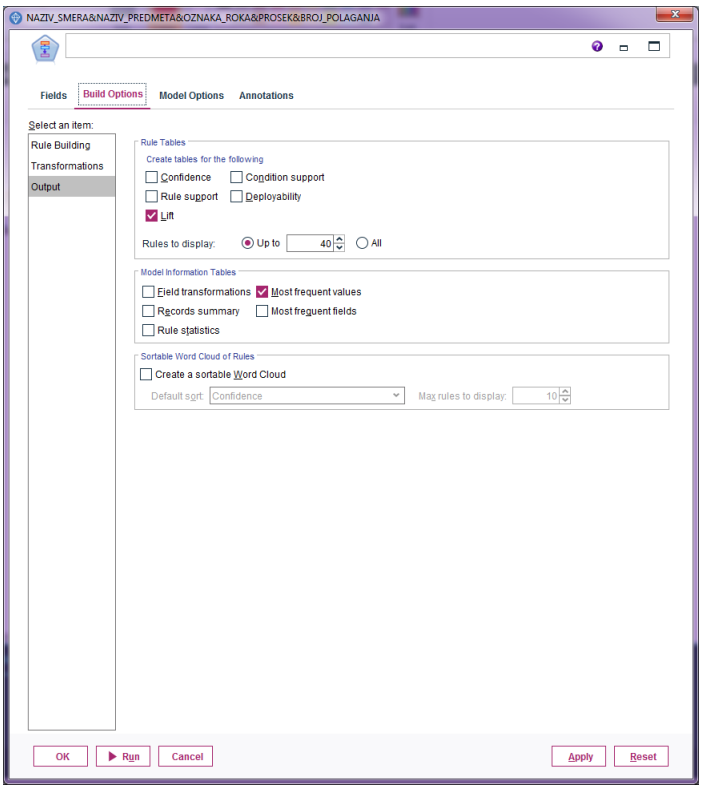
- Da atribut *GODINA_ROKA* ne bi bio korišćen pri izdvajanju pravila pridruživanja dodeljena mu je uloga *None* u čvoru **Var.** **File** za učitavanje skupa. U čvoru **Association Rules** je postavljeno da se koriste uloge atributa koje su dodeljene pre primene ovog čvora.
- Da bi numerički atributi bili podeljeni u 4 grupe, vrednost za *Number of beans* se postavlja na 4 (slika 2).
- Da bi se izdvojilo najviše 500 najboljih pravila pridruživanja prema Lift meri, vrednost parametra *Maximum number of rules* se postavlja na 500, a za parametar *Rule criterion for top N* se bira opcija Lift (slika 3).
- Prag od 5% za podršku pravila pridruživanja se postavlja preko parametra *Rule Support* (slika 3).
- Prag od 60% za pouzdanost pravila pridruživanja se postavlja preko parametra *Confidence* (slika 3).
- Da bi se izdvojila pravila pridruživanja koja su zanimljiva prema Lift meri, prag za Lift meru se postavlja na 2 pošto je moguće zadati samo celobrojnu vrednost.
- Da bi bila prikazana tabela sa pravilima pridruživanja uređenima prema Lift meri, u odeljku *Build Options*, deo *Output*, grupa *Rule Tables*, bira se opcija *Lift* (slika 4).
- Da bi bila prikazana tabela sa najčešćim stavkama u skupu, u odeljku *Build Options*, deo *Output*, grupa *Model Informations Tables*, bira se opcija *Most frequent values* (slika 4).



Slika 2: Postavljanje prametra za diskretizaciju numeričkih atributa u čvoru *Association Rules*



Slika 3: Postavljanje parametara u čvoru *Association Rules*

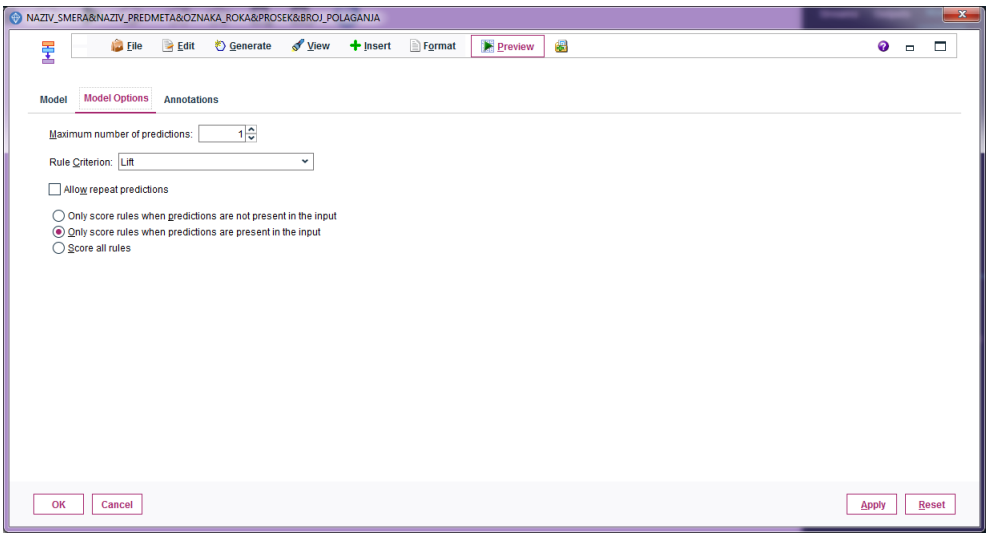


Slika 4: Postavljanje parametara u čvoru *Association Rules* za prikaz potrebnih podataka u modelu

Odgovori na pitanja se čitaju iz modela

- Među izdvojenim pravilima pridruživanja, koje pravilo je najzanimljivije?
  Prema tabeli *Most Interesting Rules by Lift* u modelu izdvojeno je 6 pravila.
  Prema Lift meri, sa vrednošću 5,85 najzanimljivije pravilo je
  OZNAKA_ROKA = okt and PROSEK $\leq$ 7.333 $\rightarrow$ 2.000 $\leq$ BROJ_POLAGANJA $<$ 3.000
  što prema znanju iz domena nije toliko zanimljivo, jer je očekivano da studenti u oktobarskom roku predmet polažu 2. ili 3. put.
  Međutim, iako ima manju vrednost Lift mere (2,87) i nižu pouzdanost, iz ugla domena, zanimljivo je i pravilo
  NAZIV_SMERA = Astrofizika $\rightarrow$ PROSEK $>$ 8.667
- Pronaći najbolje pravilo prema Lift meri za 8. transakciju u skupu kada se stavke koje su u glavi pravila javljaju u transakciji.

  Da bi se za svaku transakciju pronašlo najbolje pravilo pridruživanja koje važi za transakciju, u modelu u odeljku *Model Options* se postavljaju parametri kao na slici 5, Klikom na *Preview*, u tabeli za 8. transakciju vidi se da je najbolje pravilo sa identifikatorom 1. U tabeli *Most Interesting Rules by Lift* pravilo sa identifikatorom 1 je
  OZNAKA_ROKA = okt and PROSEK $\leq$ 7.333 $\rightarrow$ 2.000 $\leq$ BROJ_POLAGANJA $<$ 3.000, sa vrednošću za Lift meru 5,85.



Slika 5: Postavljanje parametara u odeljku *Model Options* u čvoru modela

- Koja stavka je najčešća u skupu i kolika joj je podrška?
  Prema podacima u tabeli *Information for Most Frequent Items* najčešća stavka je BROJ_POLAGANJA $\leq$ 2.000 koja se javlja u 83,19% transakcija.