Istraživanje podataka 1 - vežbe 11, 2020.

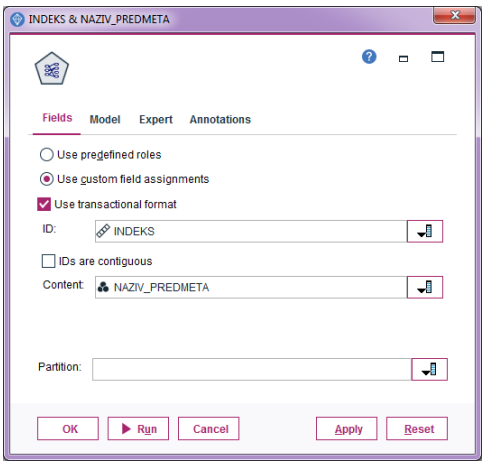
Primer 2: Primenom pravila pridruživanja proveriti da li postoji zavisnost među upisanim izbornim predmetima. Posebno napraviti model sa pravilima pridruživanja za izborne predmete studenata smera Informatika i Računarstvo i informatika.

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.

U radnom toku **pravila_pridruzivanja_izborni_predmeti.str** prvo se učitava skup podataka preko čvora *Excel*. U odeljku *Types* klikom na dugme *Read Values* učitavaju se informacije o vrednostima koje se javljaju u atributima skupa.

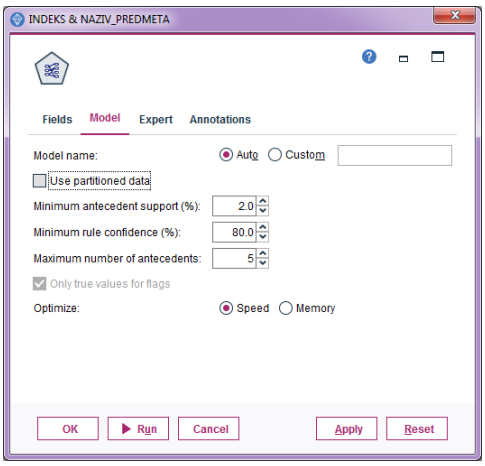
Jednu transakciju u skupu podataka predstavljaju izborni predmeti jednog studenta. Indeks studenta je identifikator transakcije, a izborni predmeti su stavke u transakcijama.

Prvo se izdvajaju pravila pridruživanja uzimajući u obzir podatke o studentima svih smerova. Ako se izuzme atribut smer, skup podataka je u transakcionom obliku. Čvor sa skupom podataka povezuje se sa čvorom *Apriori*. Preko opcija dostupnih u čvoru *Apriori*, u odeljku *Fields* navodi se da je skup u transakcionom obliku i postavljaju se podaci koji atribut sadrži identifikator transakcije, a koji atribut sadrži podatke o stavkama. (Slika 1)



Slika 1: Postavljanje parametara za transakcioni oblik podataka u čvoru *Apriori*

U odeljku *Model*, vrednost za minimalnu podršku tela se smanjuje na 2% da bi se među pravilima pridruživanja našla i pravila koja važe za izborne predmete na doktorskim studijama jer na doktorskim studijama ima značajno manje studenata u odnosu na broj studenata na osnovnim i master studijama (Slika 2).



Slika 2: Postavljanje vrednosti za minimalnu podršku tela pravila u čvoru *Apriori*

Izborom opcije *Run* pravi se model sa pravilima pridruživanja koji je u radnom toku prikazan čvorom u obliku dijamanta. Duplim klikom na model prikazuje se tabela sa izdvojenim pravilima pridruživanja. Klikom na ikonicu *Show/hide criteria menu* bira se da se u tabeli prikaže i vrednost Lift mere za svako pravilo. Preko *Sort by* padajućeg menija bira se opcija *Lift* da bi pravila pridruživanja u tabeli bila uređena prema izračunatoj vrednosti za Lift meru.

Prema Lift meri, sva pravila su zanimljiva jer najmanje zanimljivo pravilo ima vrednost 3,29 ($>$1,1). Najzanimljivije pravilo prema Lift meri ima Lift 26,171 i to pravilo je

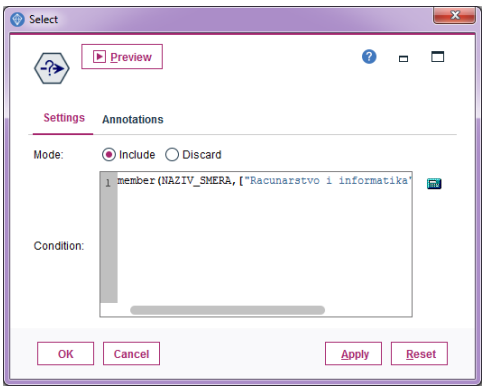
{Baze podataka-napredni koncepti = T and XML-tehnologije = T } $\rightarrow$ Masinsko ucenje = T

koje govori da studenti koji izaberu predmete *Baze podataka-napredni koncepti* i *XML-tehnologije* će verovatno izabrati i predmet *Masinsko ucenje*. Ovo su predmeti na doktorskim studijama. Zanimljivo je i pravilo Uvod u filosofiju = T $\rightarrow$ Pedagogija = T

Postoji nekoliko pravila koja su zanimljiva prema merama kvaliteta za pravila pridruživanja, a koja sadrže predmete *Taja A* i *Taja B*, npr. pravilo Taja B = T $\rightarrow$ Taja A = T. Međutim, pošto se zna da je predmet Taja A uslovni za predmet Taja B, ovo pravilo nije zanimljivo. Nakon izdvajanja zanimljivih pravila pridruživanja prema merama kvaliteta, često je potrebna i provera praktične korisnosti pravila na osnovu domenskog znanja.

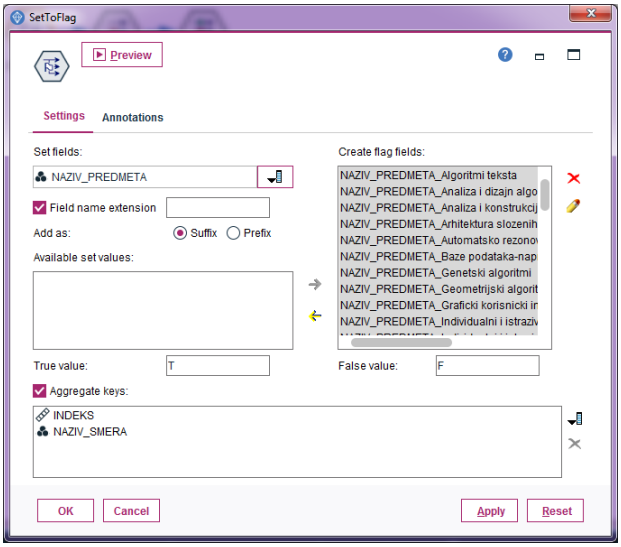
Izdvojena pravila se mogu sačuvati u HTML formatu izborom opcija *File* $\rightarrow$ *Export HTML* $\rightarrow$ *Model*.

Za izdvajanje pravila pridruživanja za izborne predmete studenata sa Informatike i Računarstva i informatike, čvor *Select* se nadovezuje na skup podataka (Slika 3).



Slika 3: Postavljanje uslova za izdvajanje samo podataka za studente sa Informatike i Računarstva i informatike

Ako je potrebno prikazati podatke o transakcijama u tabelarnom obliku, nakon izbora željenih redova, čvor *Select* se povezuje sa čvorom *Types* da bi se utvrdilo koje se sada vrednosti javljaju u atributima, tj. da bi se iz liste mogućih vrednosti za atribut *naziv_predmeta* eliminisala imena predmeta koje studenti željenih smerova nikada nisu upisali. To se postiže klikom na dugme *Clear All Values*, a zatim na dugme *Read Values*. Korišćenjem čvora *SetToFlag* vrši se binarizacija atributa *naziv_predmeta* (Slika 4).



Slika 4: Čvor *SetToFlag* za transformaciju skupa u tabelarni oblik

Pošto atributi *indeks* i *naziv_smera* nisu potrebni za dalji rad, koristi se čvor *Filter* za njihovu eliminaciju. Pomoću čvora *Types* preostalim atributima koji odgovaraju izbornim predmetima dodeljuje se uloga *Both*, čime se definiše da svaka stavka može da se pojavi u telu ili glavi pravila pri izdvajanju pravila pridruživanja.

Pravljenjem grafa povezanosti predmeta pomoću čvora *Web* uočava se jaka veza između parova predmeta: *Programske paradigme* i *Analiza i dizajn algoritama*, *Programske paradigme* i *Teorija izračunljivosti*, *Programske paradigme* i *Taja A*, *Programske paradigme* i *Taja B*, *Taja A* i *Taja B*. Ovo su bili izborni predmeti na osnovnim studijama u trenutku pravljenja baze podataka. Za ovaj skup je graf koristan za brzo uočavanje veza između predmeta na osnovnim studijama, jer je na tom nivou studija veliki broj studenata i ima manje izbornih predmeta. Nije pogodan za uočavanje jakih veza za izborne predmete na doktorskim studijama jer je mali broj studenata na tom nivou studija, a ima puno izbornih predmeta.

Radi izdvajanja pravila pridruživanja, na poslednji čvor *Types* se nadovezuje čvor *Apriori*. U odeljku *Fields* čvora *Apriori* bira se opcija *Use predefined roles* čime se pri traženju pravila pridruživanja koriste ranije dodeljene uloge atributima u čvoru *Types*. U odeljku *Model* se podrška za telo pravila pridruživanja mora smanjiti na 2% da bi se izdvojila zanimljiva pravila za studente na doktorskim studijama.

Prema Lift meri i pouzdanosti, najzanimljivije pravilo za predmete na doktorskim studijama je

Napredne arhitekture racunara → Genetski algoritmi

koje ima vrednost za Lift meru 19,5, a pouzdanost 100%.

Za osnovne studije, ako se ne uzimaju u obzir pravila sa predmetima *Taja A* i *Taja B*, ili *Strani jezik 1* i *Strani jezik 2*, izdvajaju se pravila:

Strani jezik 1 → Osnovi astronomije A
Analiza i dizajn algoritama 2 → Programske paradigme

Pravilo Strani jezik 1 → Osnovi astronomije A ima veću vrednost za Lift meru (3,228), ali nisku podršku za telo (4,272), dok pravilo Analiza i dizajn algoritama 2 → Programske paradigme ima veliku podršku za telo (20,513), ali najnižu vrednost za Lift meru (1,681).