Predlog projekta iz SIAP-a

# Definicija projekta

Ovaj dokument sadrži kratak opis teme projekta i definiciju kao i motivaciju za odabranu temu. Nakon motivacije sledi pregled vladajućih stavova i shvatanja u literaturi, zatim kratak opis skupa podataka, metodologija realizacije i evaluacije projekta, softvera koji će biti korišćen. Na samom kraju dokumenta se nalazi plan rada na projektu.

Tema projekta je analiza i istraživanje podataka koje je prikupio Centar za kontrolu i prevenciju bolesti Sjedinjenih Američkih Država u okviru nacionalnog istraživanja zdravlja članova porodice, trudnoća i prirodnog priraštaja. Fokus istraživanja biće na prikupljenim podacima o toku trudnoća i celokupnog zdravlja žena, gde ćemo se baviti eksplorativnom analizom i vizualizacijom tih podataka, kao i predikcijom nekih parametara trudnoće, kao što su trajanje ili ishod u zavisnosti od ostalih podataka vezanih za samu trudnoću i trudnice.

# Motivacija

Na nivou SAD-a godišnje se troši više od 3.2 triliona dolara na troškove zdravstva, oko 10 000 dolara po glavi stanovnika. Deo tih troškova, naravno, odlazi i na troškove zdravstvenih usluga vezanih za trudnoću. Eksplorativnom analizom ciljnog skupa podataka može se doći do informacija značajnih za bolju raspodelu novca za ulaganje u zdravstvo. Takođe, analizom podataka se može doći do vrednih saznanja koja mogu unaprediti edukativni sistem, a samim tim i poboljšati celokupno zdravlje nacije. Osim „globalnog pogleda“ na trudnoće, analizom i predikcijom ishoda, komplikacija i ostalih osobina pojedinačnih trudnoća mogu se kreirati i planirati bolji paketi zdravstvenog osiguranja.

Pregled vladajućih stavova i shvatanja u literaturi

[1] Pregnancy Outcomes by Pregnancy Intention in a Managed Care Setting, Postlethwaite, D., Armstrong, M.A., Hung, YY. et al. Matern Child Health J (2010) 14: 227. doi:10.1007/s10995-009-0446-5

<http://link.springer.com/article/10.1007%2Fs10995-009-0446-5>

### Tema rada:

Mnoge objavljene studije pokazuju da postoji korelacija između loših ishoda trudnoće i neželjenih trudnoća, tj. da je stepen loših ishoda neproporcionalno veći kod trudnoća koje su bile neželjene. Ova studija je osmišljena da posmatra ishode nadgledanih trudnoća u zavisnosti od sociodemografskog statusa žena i toga da li je trudnoća bila neželjena ili ne.

### Podaci:

Korišćeni su podaci iz 2002. godine o 1784 žene starosti od 15 do 44 godine prikupljenih na nivou Kalifornije i poređeni su sa podacima prikupljenim na nivou države, u okviru već pomenutnog Nacionalog istraživanja porodičnog rasta.

### Korišćeni algoritmi:

Korišćene su metode eksplorativne analize podataka i modeli višestruke regresije kako bi se istražili i predvideli ishodi i mogućnost javljanja neželjenih trudnoća u zavisnosti od sociodemografskog položaja žena.

### Ostvareni rezultati:

Analizom je pokazano da je 62.1% trudnoća bilo planirano, 26.4% neplanirano, a 11.4 % neželjeno. Pokazano je da su žene koje su mlade, nižeg obrazovanja, samohrane, koje konzumiraju alkohol i koje pripadaju nekoj nacionalnoj manjini, podložnije neželjenim trudnoćama. Ipak, pokazano je i da se ishodi planiranih, neplaniranih i neželjenih trudnoća statistički ne razlikuju, suprotno nekim prethodnim istraživanjima.

[2] Determinants of and Disparities in Reproductive Health Service Use Among Adolescent and Young Adult Women in the United States, 2002–2008, Kelli Stidham Hall, PhD, MSN, Caroline Moreau, MD, PhD, and James Trussell, PhD

<http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2011.300380>

### Tema rada:

Tema rada je istraživanje uticaja na stepen upotrebe zdrastvenih usluga vezanih za reproduktivno zdravlje mladih žena u SAD-u u periodu od 2002. godine do 2008. godine.

### Podaci:

Korišćeni su podaci o 4421 žene starosti od 15 do 24 godine prikupljenih u okviru više ciklusa Nacionalnog istraživanja porodičnog rasta (koje je korišćeno i u našem radu).

### Korišćeni algoritmi:

Korišćene su metode eksplorativne anlize podataka i modeli višestruke regresije kako bi se istražila upotreba zdrastvenih usluga iz ugla sociodemografskih i reproduktivnih osobina žena i na taj način istražile potencijalna unapređenja u tim zdravstvenim uslugama.

### Ostvareni rezultati:

Analizom je pokazano da je više od pola ispitanika (59%) koristilo ovaj tip zdravstvenih usluga u prethodnoj godini. Kroz regresione modele vršena je uspešna predikcija korišćenja zdravstvenih usluga u zavisnosti od starosti, obrazovanja, mesta rođenja, osiguranja, religijske pripadnosti, seksualnog iskustva, broja partnera itd. Zaključeno je da postoje nejadnakosti u korišćenju zdrastvenih usluga, pogotovo među najmlađim i socijalno najugroženijim ženama, što se može odraziti na njihovo reproduktivno zdravlje.

# Skup podataka

Za ovu temu korišćen je skup podataka prikupljen od strane Centra za kontrolu i prevenciju bolesti Sjedinjenih Američkih Država (engl. US Centers for Disease Control and Prevention –CDC) koji od 1973. godine sprovodi nacionalno istraživanje porodičnog rasta (engl. National Survey of Family Growth, u daljem tekstu NSFG). Ovo istraživanje ima za cilj da prikupi informacije o životu u porodici, sklopljenim brakovima, razvodima, trudnoćama, upotrebi kontraceptivnih sredstava, kao i zdravlju muškaraca i žena. Rezultati ovog istraživanja se koriste u svrhe planiranja zdrastvenih usluga, edukacionih programa, kao i za statistička istraživanja porodica, prirodnog priraštaja i zdravlja. Ovo istraživanje je vršeno u više ciklusa, gde ćemo u svrhe ovog rada koristiti podatke prikupljenje u šestom ciklusu, koji obuhvata period od januara 2002. do marta 2003. godine. Korišćeni podaci su dostupni na internet stranici ove studije: <http://www.cdc.gov/nchs/nsfg/nsfg_cycle6.htm> .

NSFG predstavlja studiju preseka, gde se beleže podaci jednog izabranog dela populacije u jednom trenutku vremena. Ciljna populacija ovog istraživanja su građani SAD-a između 15 i 44 godine starosti. U ovom istraživanju poseban akcenat je stavljen na različite grupe populacije, kao što su Latino-Amerikanci, Afro-Amerikanci i tinejdžeri, gde je broj ispitanika ovih grupa procentualno veći od njihove zastupljenosti u celokupnoj populaciji.

Podaci su dostupni u vidu fajlova čiji je sadržaj u ASCII formatu. Skupovi podataka su organizovani kroz parove fajlova. Prvi je „.dat“ fajl koji sadrži same podatke, organizovane u redove i kolone, gde svaki red predstavlja jednu pojavu, npr. jednu trudnoću, a drugi „.dct“ fajl predstavlja opis svake kolone, tip podatka, naziv kolone i opis vrednosti. Pošto fokus našeg istraživanja predstavlja analiza trudnoća i zdravlja žena, najviše ćemo koristiti skup podataka o trudnoćama, kao i skup podataka o ženama koje su ispitivane. U okviru skupa podataka o trudnoćama dostupni su podaci o dužini trudnoće, ishodu, redosledu trudnoće, toku, o deci rođenoj u ovoj trudnoći ili od iste majke, kao i još preko 200 drugih vrednosti koje su od manjeg ili većeg značaja za naše istraživanje. Podaci o ispitanicima sadrže podatke o starosti, trenutnom nivou zdravlja, istoriji bolesti, nivou obrazovanja i još preko 2000 drugih obeležja.

# Metodologija

Cilj našeg projekta je kreirati regresioni model putem koga ćemo vršiti predikciju rizika koje trudnoća nosi u zavisnosti od osobina same trudnoće i trudnica. Kako bismo uspešno kreirali model prvo moramo prilagoditi podatke, prilagođavanjem mernih jedinica našem metričkom sistemu, izvršiti transformaciju nenumeričkih vrednosti, izbaciti nevalidne podatke i izabrati skup podataka za obučavanje i testiranje. Nakon toga, eksplorativnom analizom ćemo pokušati da prepoznamo indikacije nekih korelacija među podacima, kao i da odgovorimo na neka opšta pitanja od interesa. Na kraju, nakon odabira nezavisnih i zavisnih promenljivih, pristupa se kreiranju regresionog modela i samoj predikciji podataka.

# Softver

Za izradu projekta koristiće se programski jezik Python, kao i bibloteke podesne za statističku analizu podataka. Biblioteka “Pandas” za prikazivanje i analizu podataka, “Numpy” za osnovne numeričke operacije, “SciPy” u svrhe naprednih izračunavanja uključujući i statistiku. “StatsModels” za regresiju i druge statističke analize, kao i “matplotlib” za vizualizaciju podataka.

# Metod evaluacije

Evaluacija regresionih rezultata će biti izvršena putem analize koeficijenta determinacije (R2) i standardne greške procene, kako na skupu za obučavanje, tako i na test skupu podataka.

# Članovi tima i plan rada

Tim čine Jelena Ilić (R1 2/2016) i Gavrilo Drljača (E2 58/2012). Plan rada na ovom projektu obuhvata sledeće bitne tačke (milestones):

* Prikupljanje podatka
* Transformacija podataka
* Eksplorativno istraživanje podataka i vizuelizacija
* Kreiranje regresionog modela
* Predikcija podataka regresijom
* Evaluacija rezultata predikcije
* Vizuelizacija dobijenih rezultata