Predlog projekta iz SIAP-a

# Definicija projekta

# Motivacija

# Relevantna literature

# Skup podataka

Za ovu temu koristili smo 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 vrešeno 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 podacija 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 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 žena čiji podaci su prikupljani. U okviru skupa podataka o trudnoćama dostupni su podaci o dužini trudnoće, ishodu, redosledu trudnoće, toku, o deci rodjenoj 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 ženskog pola sadrže podatke o starosti, trenutnom nivou zdravlja, istorji bolesti, nivou obrazovanja, pripadnosti nekoj grupi i još preko 2000 drugih obeležja.

Pre same analiza ovog skupa podataka, potrebno je izvrišiti transformaciju ovih podataka. Transformacija podataka je potrebna iz više razloga, pre svega radi eliminacije netačnih vrednosti i prilagođavanja mernih jednica našem, metričkom sistemu.

# Metodologija

Koristićemo metod regresije za predikciju 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 evauliacije

Kako bismo evaluirali rezultate naše analize, vršićemo analizu na test skupu podatka, za koji su već rezultati već poznati i uporediti ih sa dobijenim rezultatima.

# Plan

Plan rada na ovom projektu obuhvata sledeće bitne tačke (milestones):

* Prikupljanje podatka
* Transformacija podataka
* Analiza podataka upotrebom naprednih statističkih i regresivnih tehnika
* Evaluacija rezultata analize
* Vizuelizacija dobijenih rezultata

# Tim

Tim čine Jelena Ilić i Gavrilo Drljača (E2 58/2012).