Predviđanje životnog veka stanovništva

Sara Popov, IN 41/2017, popov.sara@uns.arc.rs

I. UVOD

Ovaj rad bazira se na predviđanju životnog veka stanovništva jedne države. Kako je glavni cilj velikog broja medicinskih i ostalih istraživanja detekcija problema koji narušavaju ljudsko zdravlje, pa samim tim i živote, ova analiza i predviđanje su veoma bitne. Kada se razumeju faktori, odnosno obeležja od kojih zavisi prosečni životni vek jedne populacije, moguće je sagledati mogućnosti za poboljšanje određenih društvenih faktora, kako bi predikcija bila što bolja, odnosno predviđeni životni vek stanovništva bio što veći.

Cilj analize ovih podataka jeste utvrđivanje uticaja različitih elemenata na očekivan životni vek, kao što su budžet zdravstvenih sistema, mortalitet novorođenčadi i odraslih, način života, školovanje, stepen vakcinacije i slično, za različite države. Kreirati model koji će na osnovu datih informacija moći da predvidi životni vek ljudi, a zahvaljujući kom će moći da se utvrdi u kojim državama treba povećati budžet zdravstvenog sistema, poboljšati edukaciju o zdravoj ishrani i slično.

II. BAZA PODATAKA

Baza podataka je formirana na osnovu podataka koje je objavila Svetska zdravstvena organizacija, dok su ekonomski pokazatelji objavljeni od strane Ujedinjenih Nacija. Sastoji se od 2938 uzoraka i 22 obeležja, od kojih su dva kategorička, a ostala numerička. Jedno od kategoričkih obeležja je država za koju su prikazani podaci. Istraživanje je vršeno za 193 države, u periodu između 2000. – 2005. godine i jedan uzorak predstavlja podatak o parametrima iz grupe faktora imunizacije, smrtnosti, ekonomskih faktora i socijalnih faktora za jednu određenu državu jedne određene godine.

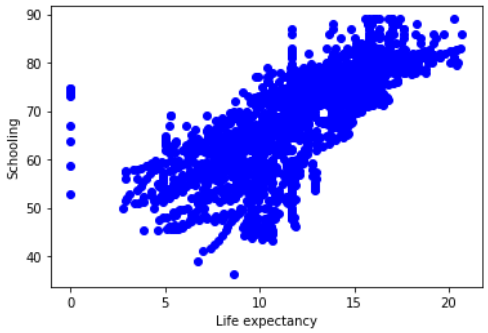
Pre rešavanja regresionih problema neopodno je detaljno analizirati bazu. Prilikom analize baze utvrđeno je da ima dosta nedostajućih vrednosti, koje su rešene na više načina. Uzorci koji imaju nedostajaću vrednost za obeležje koje se predviđa, *Life expectancy*, biće izbačeni, kako bi estimacija uspešnosti implemetiranih regresora bila tačnija. Svi uzorci kojima fali 30% ili više vrednosti obeležja takođe su izbačeni bez dopunjavanja nedostajućih vrednosti. Za sve ostale slučajeve način popunjavanja nedostajućih vrednosti vršen je kroz dodatno konsultovanje kroz pomoćnu literaturu i uočavanjem pravilnosti, odnosno paterna. Za zabeleženu potrošnju alkohola po stanovniku u litrima (obeležje *alcohol*) je karakteristično da postoji značajna razlika u stopi konzumacije alkohola u različitim državama. Na primer, Estonija ima zabeleženu potrošnju od 17.87l godišnje po glavi stanovnika, dok Avganistan ima potrošnju od 0.01l. Na osnovu toga, verodostojniji podaci dobijaju se ukoliko se nedostajuće vrednosti za ovo obeležje umesto popunjavanjem medijanom obeležja od svih uzoraka popunjavaju medijanom uzoraka te države. Isti postupak bi trebalo da se ponovi za obeležje Hepatitis B, međutim ovo obeležje za 9 država ima sve nedostajuće vrednosti. Kako se stopa vakcinisanih protiv Hepatitisa B značajno razlikuje od toga da li je država razvijena, 88.041% ili je u razvoju, 79.782%, nedostajuće vrednosti popunjene su medijanom uzoraka istog statusa države. Isti princip primenjen je na nedostajućim vrednostima za obeležje *Polio* (imunizacija protiv dečije paralize), za obeležje *Diphtheria tetanus toxoid* (imunizacija protiv tetanusa), za obeležje *Income composition of resources* (indeks humanog razvoja u smislu prihoda) i za obeležje *Schooling*, takođe zbog uočenih pravilnosti. Nedostajuće vrednosti za obeležja *BMI* (indeks telesne mase), *Thinness 5-9 years* (Rasprostranjenost mršavosti kod dece i adolescenata u uzrastu od 5 do 9 godina), *Thinness 10-19 years* (Rasprostranjenost mršavosti kod dece i adolescenata u uzrastu od 10 do 19 godina), popunjene su medijanom tog obeležja nad celokupnom bazom. Nedostajuće vrednosti obeležja *Total expenditure* (procenat rashoda koji je potrošen na zdravstvenu zaštitu) i obeležja *GDP* (Bruto domaći proizvod) popunjene su medijanom vrednosti obeležja te zemlje, a za one zemlje koje nemaju ni jednu popunjenu vrednost za to obeležje uzeta je medijana ostalih država sa istim statusom. Kod obeležja *Population* nedostajuće vrednosti popunjene su takođe medijanom ostalih uzoraka koje pripadaju toj državi, a za one zemlje koje nemaju ni jednu vrednost ovog obeležja popunjene su medijanom obeležja celokupne baze podataka.

III. ANALIZA OBELEŽJA I NJIHOVIH KORELACIJA

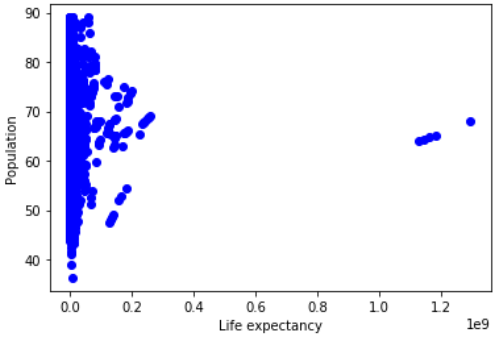
U matrici korelacije se jasno vidi da je oeležje koje je izlaz regresije u najvećoj korelaciji sa obeležjem *Schooling* (slika 1), u najmanjoj korelaciji sa obeležjem *Population* (slika 2), a najveći stepen negativne korelacije javlja se sa obeležjem *Adult Mortality* (slika 3).

Obeležja *Country* i *State* su kategorička obeležja i kao takva ne mogu se koristiti za potrebe linearne regresije, nego su njihove vrednosti pretvorene u numeričke.

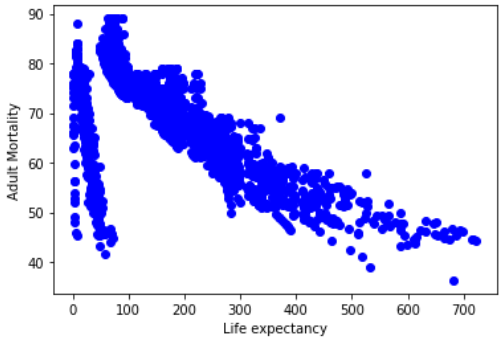
U kodu je implementirana posebne funkcija pod nazivom *dataSetAnalises* koja daje uvid u statističke veličine i raspodele. Ovde će biti prokomentarisana raspodela obeležja koje se predviđa (slika 4). Dinamički opseg *Life expectancy* obeležja je 53, najniže predviđeni životni vek je 36 godina, a najviši 89, dok se 50 % uzoraka nalazi u opsegu predviđene 72 godine.



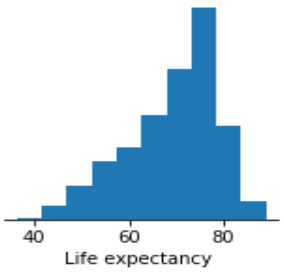
Slika 1. Prikaz korelacije obeležja *Schooling* i *Life expectancy*



Slika 2. Prikaz korelacije obeležja *Population* i *Life expectancy*



Slika 3. Prikaz korelacije obeležja *Adult Mortality* i *Life expectancy*



Slika 4. Prikaz raspodele obeležja *Life expectancy*

IV.PREDVIĐANJE ŽIVOTNOG VEKA STANOVNIŠTVA

Nad prethodno opisanim podacima je urađena standardizacija, kako bi se ubrzala konvergencija metode opadanja gradijenta. Skaliranje je izvršeno na isti opseg vrednosti, kako bi bile sprečene da se potencijalno odluči o izlaznoj vrednosti na osnovu manjeg broja obeležja, koje variraju u većem opsegu vrednosti. Baza podataka nije inicijalno podeljena na trening i test skup, a tokom kreiranja regresivnih rešenja koristiće se metoda unakrsne validacije. Unakrsna validacija odabrana je kako bi se prevazišao problem korišćenja nekih uzoraka samo u svrhu testiranja, a posledica toga je smanjenje natprilagođenja.

Za rešavanje regresionog problema nad datim skupom podataka isprobane su više različitih metoda, koje su i predstavljene u kodu. Odabrane su tri, koje će i u okviru ovog pasusa biti objašnjene.

Prva metoda koja je korišćena za predviđanje obeležja *Life expectancy* je linearna regresija sa hipotezom y=b0+b1x1+...+bnxn, odnosno kada na izlaz modela utiču samo ostala obeležja koja su uzeta za predviđanje i prethodno pripremljeni podaci. Primenom ove hipoteze dobija se model kod kojeg je prosečni R2 = 0.920, prosečni MAE = 1.87, a prosečni MSE = 6.74 (prosečni se odnosi na prosek od 10 izvršenih iteracija primenom metode unakrsne validacije).

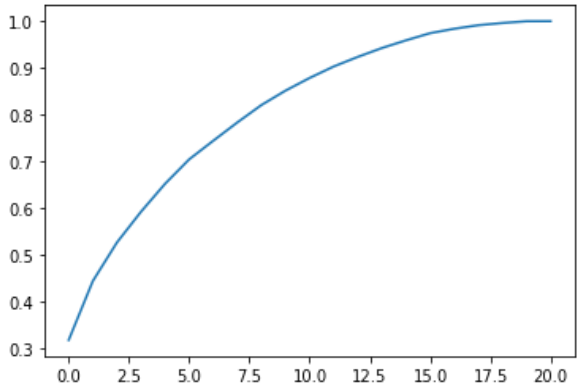
Sledeća odabrana metoda je stablo odluke. Funkcionisanje ovog regresora zasniva se na grananju stabla, gde koren stabla predstavlja čvor koji sadrži skup svih uzoraka i od njega se vrši sukcesvina particija skupa uzoraka na dva disjunktna podskupa. Odluka o tome po kom obeležju će u datom koraku biti izvršena particija ekvivalentna je pitanju koje je obeležje najznačajnije u tom trenutku. Karakteristično za rešavanje regresionih problema putem stabla odluke je da se prilikom obuke modela formira stablo odluke, a regresija se vrši tako što se novi uzorak propusti kroz stablo i na osnovu toga kom listu pripada dodeljuje mu se prosek vrednosti uzoraka tog lista. Utvrđeno je da su optimalni parametri za ovu metodu da je kriterijum srednja kvadratna greška (*Mean squared error MSE*), maksimalna dubina stabla (*max_depth*) da je 10, minimalni procenat uzoraka koji čvor sme da ima a da je dozvoljeno da se deli (*min_samples_split*) je 0.05. Parametar kriterijum meri kvalitet podele. Kroz testiranje obučenog stabla odluke pokazano je da ovakav način implementacije modela regresije daje bolje rezultate od klasične linearne regresije. Prosečni R2 je 0.936, što je svakako bolje od R2 linearne regresije, prosečni MAE = 1.56, a prosečni MSE = 5.47. Prednost rešavanja regresionog problema pomoću stabla odluke je mogućnost jednostavne vizualizacije (dato u kodu) i interpretacije. Velika dimenzionalnost ne predstavlja problem kod ove metode, a lako uočavanje najvažnijih obeležja predstavlja takođe prednost. Problem preobučavanja je negativna tendencija stabla odluke. Metoda slučajne šume u određenoj meri rešava pomenute probleme obučavajući mnoštvo stabala odluke i donošenje krajnje odluke glasanjem. Ako je dovoljno velik skup podataka metoda slučajne šume se pokazala kao bolja opcija od jednog stabla odluke i zato je implementirana ta metoda. R2 predviđenjem metodom slučajne šume je 0.956, što je bolje nego kod običnog stabla odluke, prosečni MAE = 1.22, a prosečni MSE = 3.77.

Sledeći regresor koji će biti korišćen je SVR, odnosno mašina na bazi vektora nosača prilagođena rešavanju regresionih problema. Parametri koji bi trebalo da budu odbrani su koji će se kernel koristiti, odnosno regularizacioni parametar iz ciljne funkcije (C), nezavisan član za polinomijalni i sigmoidni kernel (coef0), stepen polinomijalnog kernela (degree). Čak i najbolje odabrana kombinacija parametara, a ona je da je: C = 0, coef0 = 0.001, kernel = ‘poly’, daje loše rezultate, odnosno R2 je u minusu što govori do pojave probučavanja, prosečni MAE = 17.44, a prosečni MSE = 59206.72. Potrebno je naglasiti da nije bilo mogućnosti da se ispitaju neke određene kombinacije parametra, iz određenog nepoznatog razloga to ispitivanje je trajalo par sati, pa je zbog toga proces prekinut.

V. PCA

Za obuku postoji maksimalni broj obeležja koji, kada se prekorači, rezultuje opadanju performansi algoritma. Sakupljanje podataka i proširivanjem baze podataka često je nemoguć proces, jedan od razloga je što je to skup proces. Drugi način za poboljšanje performansi, potencijalno ubrzanje algoritma i pojednostavljenja modela je smanjenje dimenzionalnosti. Kroz ovo poglavlje biće objašnjena redukcija dimenzionalnosti, koja podrazumeva formiranje manjeg skupa svih raspoloživih obeležja kombinovanjem postojećih. Glavni cilj redukcije obeležja je da od zadatog skupa obeležja stvori novi skup obeležjima, takvim da neće postojati redudantni podaci. Korišćena je tehnika razlaganja na glavne komponente (*Principal Component Analysis PCA*), ova tehnika redukcije spada u metodu nenadgledanog učenja i koristi kriterijum reprezentacije. Ideja ove metode je da se informacija koju obeležja prenose zapravo krije u varijansi podataka i zato je potreban proces pronalaženja najveće varijanse podataka. Cilj PCA algoritma je da očuva varijansu sadržanu u podacima, a funkcioniše tako što pronađe pravac najbržeg rasipanja i taj pravac predstavlja prvu PCA komponentu, naredna PCA komponenta će takođe predstavljati linearnu kombinaciju određenih obeležja, čija je zadatak takođe da opiše što više varijanse. Poznato je da će druga PCA komponenta opisati manje podataka od prve i potrebno je da bude normalnu u odnosu na nju, odnosno PCA komponente su međusobno nezavisne.

Prvi korak prilikom korišćenja PCA metode je standardizacija obeležja, odnosno centriranje uzoraka. Parametar koji je potrebno podesiti je koja će biti dimenzionalnost novog prostora. Kako ne postoji ustaljeno pravilo koje daje uvek dobre rezultate o tome koliki procenat varijanse je potrebno zadržati, korisno je pogledati kako se za ove podatke menja udeo objašnjene varijanse (Y osa) kroz promenu broja PCA komponenti (X osa).



Slika 5. Broj PCA komponenti

Parametaru koji određuje koliko komponenti će predstavljati novi prostor (*n_components*) zadata je vrednost 0.9 i to znači da će biti kreirano onoliko komponenti koliko je dovoljno da se objasni 90% varijanse.

VI.USPEŠNOST REGRESORA NAKON REDUKCIJE DIMENZIONALNOSTI

Kao što je već napomenuto, redukcija dimenzionalnosti pomoću PCA algoritma ne garantuje ostvarivanje boljih performansi algoritma u poređenju sa performansama postignutih nad originalnim skupom obeležja. Upravo zbog ovog razloga su pokrenute sve tri metode rešavanja regresionog problema ponovo i za Linearnu regresiju, kao i za Metodu stabla odluke, Slučajne šume i SVR. U nastavku poglavlja biće predstavljeni njihovi rezultati, kao i kompentari potencijalnog nastanka odstupanja. Bitno je napomenuti da je unakrsna validacija koja je rađena pre PCA algoritma rađena na celim skupom podataka, a da je nakon PCA algoritma rađena na odvojenom trening skupu.

U nastavku poglavlja biće predstavljeni njihovi rezultati, kao i komentari potencijalnog nastanka odstupanja. Bitno je napomenuti da je unakrsna validacija koja je rađena pre PCA algoritma rađena na celim skupom podataka, a da je nakon PCA algoritma rađena na odvojenom trening skupu.

	Pre redukcije dimenzionalnosti	Posle redukcije dimenzionalnosti
R2	0.92	0.857
MSE	6.745	12.354
MAE	1.87	2.58

Tabela 1. Mere uspešnosti linearne regresije

	Pre redukcije dimenzionalnosti	Posle redukcije dimenzionalnosti
R2	0.936	0.834
MSE	5.477	14.286
MAE	1.560	2.744

Tabela 2. Mere uspešnosti stabla odluke

	Pre redukcije dimenzionalnosti	Posle redukcije dimenzionalnosti
R2	0.955	0.895
MSE	3.774	9.036
MAE	1.226	2.117

Tabela 3. Mere uspešnosti metode slučajne šume

	Pre redukcije dimenzionalnosti	Posle redukcije dimenzionalnosti
R2	-0.577	0.708
MSE	97.660	25.096
MAE	3.478	3.440

Tabela 4. Mere uspešnosti SVR metode

Iz prethodnih tabela primetno je da se nakon redukcije dimenzionalnosti tačnost kod linearne regresije, stabla odluke i metode slučajne šume smanjuje tačnost i da njima nije „prijala“ ova promena, međutim potrebno je uočiti veliku promenu kod SVR metode. Pre redukcije dimenzionalnosti ova metoda davala je najlošije rezultate, a narušena tačnost bila je posledica preobučavanja. Međutim nakon redukcije dimenzionalnosti R2 više nije negativan, i dalje nije dovoljno dobar kao što je R2 kod metode slučajne šume na primer, ali svakako je značajno bolji nego pre rudukcije dimenzionalnosti i sada problem probučavanja nije više toliko izražen.

Na osnovu predstavljenog moguće je zaključiti da ovakvu redukciju dimenzionalnosti nije potrebno raditi prilikom primene metode linearne regresije, stabla odluke i slučajne šume, tačnije nije ni poželjno, jer takva obeležja daju lošije podatke od prvobitnih, ali ukoliko je potrebno primeniti metodu SVR svakako je poželjno i potrebno. U impplementaciji ovog rada nije mereno i upoređivano vreme

izvršetka obučavanja pre i posle redukcije dimenzionalnosti, imalo bi smisla da je to vreme različito jer je različit broj obeležja, odnosno pre redukcije se obučava kroz 22 obeležja, a nakon kroz 12, ali ovo intuitavno razmišljanje nije pokazana u ovom radu.

VII. ZAKLJUČAK

Predikcija životnog veka određenog stanovništva predstavlja veoma bitnu analizu, jer na osnovu nje se stiče znanje o potencijalnim problematičnim segmentima i uviđaju se određene pravilnosti koje pri implementaciji potencijalno mogu da povećaju predviđen životni vek društva, što predstavlja krajnji cilj većine medicinskih i ekoloških istraživanja.

Kroz ovaj rad pokazano je da metoda slučajne šume daje najbolje rezultate kada se poredi sa metodom linearne regresije, stabla odluke i metodom na bazi vektora nosača za regresione probleme. Takođe, nakon odabira optimalnih parametra i obukom kroz metodu unakrsne validacije urađena je i redukcija dimenzionalnosti, kroz PCA algoritam. Pokazano je da tako obrađeni podaci daju značajno bolje rezultate za SVR metode, dok se kod ostalih mere uspešnosti smanjuju. Na osnovu toga, najbolje bi bilo za ovaj skup obeležja koristiti metodu slučajne šume na originalnom skupu podataka, jer obeležja nema relativno puno, u odnosu na neke ostale skupove podataka, pa takvo obučavanje nije preterano zahtevno, a daje najbolju meru uspešnosti u poređenju sa ostalim ispitanim metodama i parametrima.

VIII. KORIŠĆENA LITERATURA

[1] „Praktikum iz mašinskog učenja“, Tijana Nosek, Branko Brkljač, Danica Despotović, Milan Sečujski, Tatjana Lončar-Turukalo

[2] Education Index https://en.wikipedia.org/wiki/Education_Index

[3] Global Viral Hepatitis https://www.cdc.gov/hepatitis/global/index.htm

[4]List of countries by mass index https://en.wikipedia.org/wiki/List_of_countries_by_body_mass_index

[5]Human development index https://ourworldindata.org/human-development-index