Predviđanje životnog veka

Jovana Bogojević, IN32/2017, bogojevic.in32.2017@uns.ac.rs

I UVOD

U izveštaju predstavljeno je rešavanje regresionog problema, koji je primer problema nadgledanog učenja, kod kog je izlazna promenljiva kontinualnog tipa i odnosi se na očekivani životni vek ljudi.

Rezultat je formiran modela regresije koji na osnovu dostupnih obeležja iz baze predviđa životni vek ljudi zahvaljujući kom se može utvrditi povezanost između različitih faktora, poput izdvojenog budžeta za zdravstveni sistem neke države, mortaliteta ili stepena vakcinacije i njihov uticaj na prosečni životni vek stanovnika.

II OPIS BAZE PODATAKA

Baza je formirana na osnovu podataka objavljenih od strane Svetske zdravstvene organizacije, a ekonomski pokazatelji objavljeni su od strane Ujedinjenih Nacija. U bazi je sadržano 2938 uzoraka i 22 obeležja.

Obeležja koja opisuju uzorke odnose se na demografske pokazatelje poput broja stanovnika, stope smrtnosti odraslih oba pola na 1000 stanovnika, broj umrle odojčadi na 1000 živorođenih, broj smrtnih slučajeva mlađih od 5 godina na 1000 stanovnika, kao i na ostale statističke podatke – prosečni indeks telesne mase celokupne populacije (BMI), rasprostranjenost mršavosti kod dece za uzrast od 5 do 9 godina izraženo u procentima, kao i rasprostranjenost mršavosti kod dece i adolescenata za uzrast od 10 do 19 godina, broj prijavljenih slučajeva malih boginja na 1000 stanovnika i broj umrlih na 1000 živorođenih prouzrokovano virusom HIV. Dati su i podaci o pokrivenosti imunizacijom protiv Hepatitisa B, poliovirusa koji izaziva dečiju paralizu, tetanusa i pertusisa. Prikazane su i vrednosti ekonomskih pokazatelja - bruto domaćeg proizvoda i izdataka države za zdravstvo, zatim broj godina školovanja kao i zabeležena potrošnja alkohola po stanovniku starijih od 15 godina. Jedan uzorak u bazi predstavlja vrednosti svih navedenih obeležja za 193 države sveta sa statusom razvijene države ili države u razvoju od 2000. do 2015. godine.

III KOREKCIJE U BAZI PODATAKA

Pod korekcijama u bazi podrazumeva se rešavanje problema nedostajućih vrednosti.

Obeležja čije su vrednosti poznate za svaki uzorak su: *country, infant deaths, percentage expenditure, under-five deaths i HIV/AIDS.*

U *TABELI 1.* prikazan je procenat nedostajućih vrednosti za obeležja koja ih poseduju.

Naziv atributa	Br. nan vrednosti	%
Life expectancy	10	0.34
Adult mortality	10	0.34
Alcohol	194	6.6
Hepatitis B	553	18.82
BMI	34	1.16
Polio	19	0.65
Total expenditure	226	7.7
Diptheria	19	0.65
GDP	448	15.25
Population	652	22.2
thinness 1-19 years	34	1.16
thinness 5-9 years	34	1.16
Income composition	167	5.64
Schooling	163	5.54

REŠAVANJE PROBLEMA NEDOSTAJUĆIH VREDNOSTI

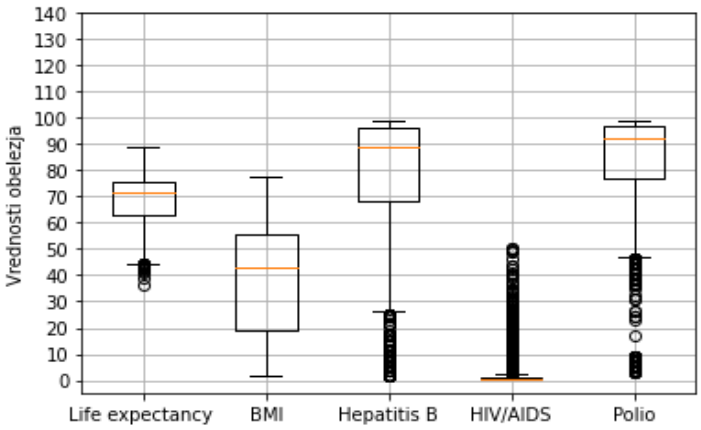
Jedan od načina popunjavanja nedostajućih vrednosti nekog od obeležja u uzorku jeste prosleđivanje poslednje validne vrednosti iz prethodnog reda napred. Ovakav pristup primenjen je za popunjavanje nedostajućih vrednosti kod obeležja *Hepatitis B, GDP, Population* (delimično) i *Diphtheria.* Na sličan način popunjene su nedostajuće vrednosti obeležja *Alcohol i Total expenditure* gde su prve validne vrednosti prosleđene red pre u kom nedostaje vrednost. Međutim, primenom prethodnih metoda rešen je problem samo jednog dela nedostajućih vrednosti. Budući da su u bazi dati podaci o svakoj državi od 2000. do 2015. godine, postoje uzorci koji se odnose na istu državu koji nemaju vrednosti nekog od obeležja za sve godine. Na primer, države Novi Zeland, Saudijska Arabija ili Libija nemaju podatke o populaciji ni za jednu godinu. Ovaj problem se ne može rešiti prostim prosleđivanjem poslednjih validnih vrednosti, jer bi onda država imala istu populaciju kao država pre nje, što nije ispravno. Takođe, ukoliko se izostavi u potpunosti svaka država koja nema sve podatke, obim podataka u bazi bi se znatno smanjio, što nije poželjno. Iz tog razloga, nedostajuće vrednosti obeležja koje se odnosi na populaciju ručno su popunjeni prosečnom vrednosti populacije za tu državu od 2000. do 2015. godine vrednostima preuzetih sa sajta koji sadrži podatke o populaciji svih zemalja sveta. Iz baze su izostavljene i države koje imaju podatke samo za jednu godinu kao i one koje nemaju nijedan podatak o obeležju *Hepatitis B.*

Kada su u pitanju kategorička obeležja, obeležje koje se odnosi na status države , u smislu da li je država u razvoju ili je ekonomski razvijena, vrednosti tog obeležja pretvorene su

u numeričke vrednosti za svaku državu, pa je broj uzoraka nakon primenjenih korekcija u bazi 2368.

III ANALIZA BAZE PODATAKA

U bazi su predstavljena merenja od 2000. do 2015. godine za države sa svih kontinenata. Analizom statističkih parametara primećeno je prisustvo netipičnih vrednosti, odnosno *outlier*-a. Na primer, za obeležje *infant deaths* najveća izmerena vrednost je 1800 što predstavlja grešku, jer vrednost ovog obeležja ukazuje na broj umrle odojčadi na 1000 živorođenih. S obzirom na da postoji još uzoraka kod kojih vrednost ovog obeležja prelazi 1000, izvršena je zamena vrednosti koje previše odstupaju posmatrajući statističke podatke datog obeležja. Na sličan način rešene su netipične vrednosti i kod ostalih obeležja koja su ih posedovala kao na primer kod obeležja: *percentage expenditure, measles i under-five deaths.*

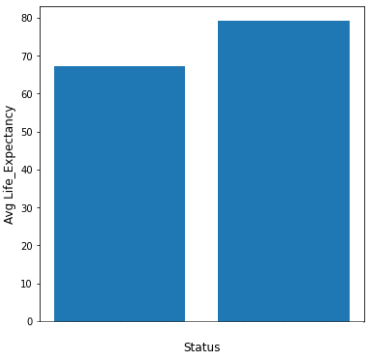


SLIKA 1: Vrednosti obeležja Life expectancy, BMI, Hepatitis B, HIV/AIDS, Polio prikazane putem boxplot-ova

OČEKIVANA ŽIVOTNA STAROST

Obeležje *Life expectancy*, odnosno očekivana životna starost odnosi se na očekivano trajanje životnog veka ljudi u godinama.

Na osnovu *SLIKE 2* (na sledećoj strani) može se videti prvih 10 država sa najvećom prosečnom životnom starosti stanovništva: Švedska, Francuska, Italija, Australija, Kanada, Austrija, Singapur, Novi Zeland, Izrael kao i Grčka, sa prosečnom vrednosti oko 80 godina. Država sa najmanjom prosečnom životnom starosti se nalazi u zapadnoj Africi – Sijera Leona sa prosekom od oko 45 godina. Za Republiku Srbiju zabeležena je prosečna starost od oko 73 godine, a za ostale države iz regiona: Bosna i Hercegovina – 75 godina, Hrvatska – 76 godina itd..

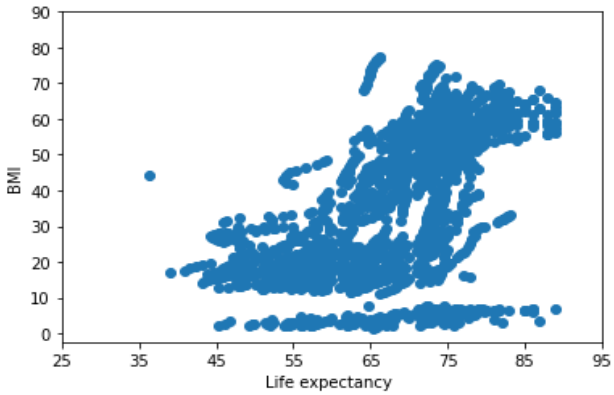


Na osnovu *SLIKE 2 i 3* može se uvideti da države sa statusom „u razvoju“ imaju manju prosečnu starost stanovništva u odnosu na države koje su ekonomski razvijenije države.

SLIKA 3

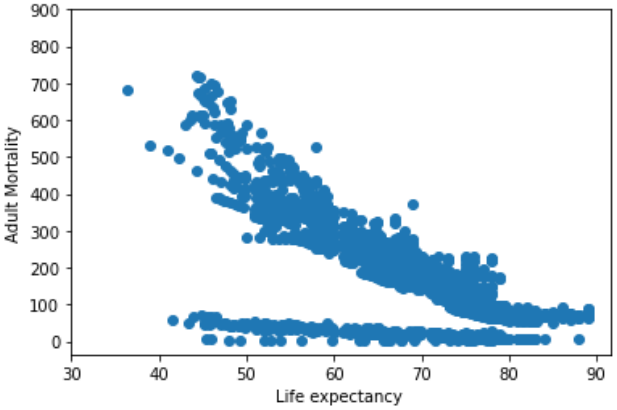
ZAVISNOSTI IZMEĐU ATRIBUTA

Sa *SLIKE 4*. može se konstatovati pozitivna korelacija (u iznosu od 0.58) između očekivane životne starosti i BMI indeksa. Ova pozitivna korelacija pokazuje da zdravije prehrambene navike, način života i vežbanje imaju dobar uticaj na kvalitet i dužinu života.

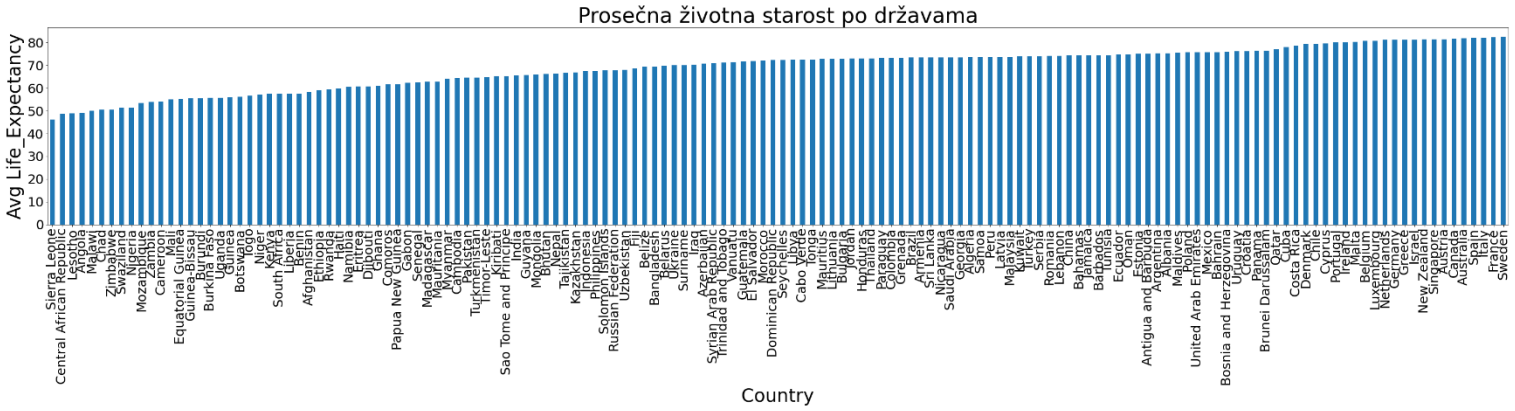


SLIKA 4: *Korelacija između obeležja Life expectancy i BMI*

Negativna korelacija najizraženija je između očekivanog životnog veka i stope smrtnosti odraslih oba pola (-0.68). Očigledno je da država sa većom stopom smrtnosti odraslih ima nižu očekivanu dužinu životnog veka i obrnuto.



SLIKA 5: Korelacija između obeležja Life expectancy i Adult mortality



SLIKA 2: Prosečna životna starost po državama

IV PREDVIĐANJE OČEKIVANE ŽIVOTNE STAROSTI

S obzirom na to da je predviđanje izlazne promenljive u ovom slučaju regresioni problem, primenjeni su algoritmi namenjeni toj vrsti problema.

Pre primene algoritama, skup podataka podeljen je na trening (na osnovu kojih se predviđaju nove vrednosti – 90% podataka) i test skup (podaci koje istrenirani model prvi put vidi – preostalih 10%). Takođe, primenjena je i standardizacija obeležja kako bi se obeležja skalirala na isti opseg vrednosti sa krajnjim ciljem ubrzavanja obuke.

MODEL LINEARNE REGRESIJE

Kao prvi model, odabrana je *Lasso* regresija – regularizaciona tehnika koja ima za cilj ograničenje procene koeficijenata koji se nalaze uz promenljive i postizanje kompromisa između pristrasnosti i varijanse. Da bi se odredila optimalna vrednost regularizacionog parametra *alpha* upotrebljena je unakrsna validacija. Unakrsna validacija služi za prevazilaženje problema korišćenja nekih uzoraka samo u svrhu testiranja, tako što se svaki od uzoraka više puta koristi za treniranje, a jednom za testiranje. Metod unakrsne validacije biće iskorišćen prilikom obuke svih modela u ovom radu.

Za vrednost regularizacionog parametra *alpha* = 0.5 dobijeni su sledeći rezultati: R^2 skor iznosi 0.85, prilagođeni R^2 skor 0.83, a srednja apsolutna greška +-3.10 godina. Dobijeni rezultati su zadovoljavajući, ali verovatno mogu biti poboljšani.

Pored *Lasso* regresije, upotrebljena je i *Ridge* regresija. Unakrsnom validacijom kao optimalni parametar *aplha* dobijena je vrednost 0.5, kao kod prethodnog modela. Međutim rezultati kod ovog regresionog modela su znatno bolji nego prethodni:

R^2 skor	prilagođeni R^2 skor	mae
0.97	0.96	1.41

kNN REGRESOR

kNN algoritam (engl. *k nearest neighbors*) se koristi kako za klasifikacione probleme, tako i za regresione. U slučaju regresionog problema, kNN regresor se obučava na različitim skupovima za obuku, a vrednost nepoznate vrednosti dobija se usrednjavanjem rezultata. To praktično podrazumeva da se nepoznatom uzorku dodeljuje vrednost za očekivanu životnu starost na osnovu toga koliko podseća na tačke u skupu za trening.

Elementi koji su neophodni da bi se primenio kNN metod su skup uzoraka za obuku, celobrojni parametar *k* i metrika. Parametar *k* označava broj najbližih suseda koji se uzima u obzir pri odlučivanju, a na osnovu odabrane metrike meri se udaljenost između novog uzorka i svakog uzorka iz skupa za obuku. Neophodno je da se odaberu optimalne vrednosti za metriku i parametar *k* kako bi se dobilo što bolje ponašanje regresora.

Prvi korak je računanje udaljenosti. Na primer, Euklidsko rastojanje računa se kao kvadratni koren sume kvadratnih razlika između nove tačke i postojeće tačke. U suštini, izbor metrike zavisi od tipa obeležja. Sledeći korak je određivanje parametra *k*, koji određuje broj suseda koje uzimamo u obzir na osnovu kojih se dodeljuju vrednosti novoj opservaciji. Pored metrike i parametra *k*, posmatra se i vrednost parametra *weights* koji služi za dodeljivanje značaja svakom od *k* najbližih suseda u odlučivanju u zavisnosti od toga da li svi susedi imaju jednak značaj ili je obrnuto proporcionalan njihovoj udaljenosti od novog uzorka.

Primenom unakrsne validacije, kombinacija parametara koja se pokazala kao optimalna jeste: metrika – *minkovski* (predstavlja uopštenje euklidskog rastojanja, dodaje slobodan parametar *p*), broj suseda – 3 i rastojanje – *distance*. Rezultati za model obučen nad ovakvim parametrima:

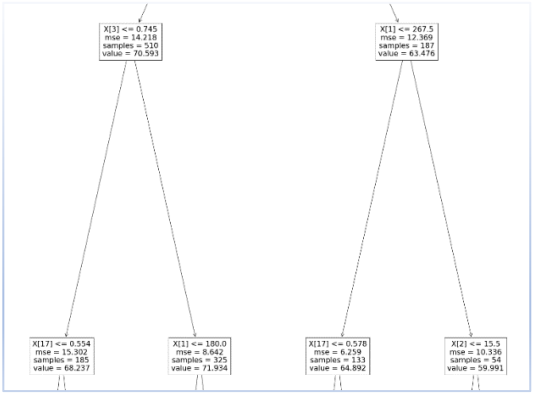
R^2 skor	prilagođeni R^2 skor	mae
0.97	0.97	1.03

Primećuje se poboljšanje greške u odnosu na prošli model.

STABLA ODLUKE

Metoda nadgledanog učenja koja se često koristi. Koren stabla odluke sadrži skup svih uzoraka od koga se stablo dalje grana. Ovaj metod donosi procenu postavljanjem niza pitanja, a svako pitanje sužava moguće vrednosti sve dok model ne postane dovoljno siguran da može napraviti jedno predviđanje. Formiranje stabla odluke podrazumeva obuku modela, a regresija se vrši propuštanjem uzorka kroz stablo i proceni u kom listu će završiti. Kriterijum podele skupa uzoraka kod regresionih problema je srednja kvadratna ili srednja apsolutna greška. Kriterijum zaustavljanja grananja čvorova može biti dobijanje čistog čvora ili da se zaustavi na maksimalnoj zadatoj dubini stabla.

Kombinacija parametara za obuku modela stabla odluke koja daje najbolji rezultat je: za kriterijum podele – srednja kvadratna greška, *max-depth =15* (predstavlja maksimalnu dubinu stabla) i *min_samples_split* = 0.01 (predstavlja udeo broja uzoraka u čvoru da bi bilo dozvoljeno njegovo deljenje).



SLIKA 6: *Deo stabla odluke*

Na osnovu odabranih parametara, dobijeni su sledeći rezultati za mere uspešnosti modela: R^2 skor – 0.94, prilagođeni R^2 skor – 0.93, srednja kvadratna greška – +-1.53. Dobijena greška manja je od dva prethodna modela koja su analizirana.

V SMANJENJE DIMENZIONALNOSTI - PCA

PCA (engl. *Principal Component Analysis*) je metoda nenadgledanog učenja koja se često koristi za smanjenje dimenzionalnosti velikih skupova podataka. Cilj je da se uzorci iz visokodimenzionalnog prostora predstave u prostoru sa manjim brojem dimenzija vodeći računa da se sačuva što više informacija o polaznom skupu podataka, budući da je manje skupove podataka lakše istraživati i vizualizovati. Dakle, PCA je statistička procedura kojom se skup od eventualno korelisanih atributa, preslikava u skup nekorelisanih atributa. Svaki od atributa iz novog skupa se naziva glavna komponenta, a svi oni su *linearno nekorelisani*. Prva glavna komponenta se bira tako da ima najveću moguću varijansu. Svaka naredna glavna komponenta se bira tako da je normalna na sve prethodne i da, takođe, ima trenutnu najveću varijansu.

Primena PCA metode na bazu koja je tema ovog izveštaja podrazumevala je da su obeležja prethodno standardizovana, a cilj je bio da se zadrže komponente koje obuhvataju 90% ukupne varijanse. Nakon redukcije, redukovani prostor ima dimenziju 135.

Nakon redukcije dimenzionalnosti prostora, izvršena je obuka prethodno opisanih modela (linearna regresija, knn regresor i stablo odluke) da bi se mogle uporediti performanse obučenih modela pre i posle redukcije.

	R^2 skor	prilagođeni R^2 skor	mae
Lasso	0.84	0.83	3.05
Ridge	0.91	0.91	2.18
kNN	0.97	0.97	1.04
DTR	0.95	0.95	1.52

TABELA 2: *Mere uspešnosti modela nakon redukcije*

Na osnovu dobijenih rezultata može se uvideti da su za model *Lasso* regresije i *kNN* klasifikatora mere uspešnosti ostale identične, kao pre redukcije. Kod modela *Ridge* regresija došlo je do pogoršanja srednje apsolutne greške – za 0.77, a kod *DTR* došlo je do blagog poboljšanja R^2 skora i R^2 prilagođenog skora.

Zaključak je da se model kNN regresije pokazao kao najbolji, kako za prostor pre primene redukcije, tako i za redukovani prostor obeležja.

VI ZAKLJUČAK

Za početak, u bazi je bilo prisutno mnogo nedostajućih vrednosti, te je bio izazov rešiti taj problem. Takođe je bilo mnogo grešaka u smislu vrednosti podataka za neka obeležja na koje je skrenuta pažnja na prethodnim stranama izveštaja. Sam cilj analize ove baze podataka je utvrđivanje uticaja različitih faktora na očekivani životni vek ljudi. Moglo se primetiti da prosečan životni vek ljudi zavisi od same ekonomske razvijenosti države, gde viši standardi života očekivano pozitivno utiču na životni vek – na primer, korelacija između broja godina školovanja (obeležje *schooling*) i životnog veka je pozitivna, kao i sam pozitivan uticaj stepena imunizacije stanovništva. Negativna korelacija primećena je između broja stanovništva i očekivane životne starosti. Kako se stanovništvo povećava, očekivano trajanje života se smanjuje, a moguć razlog za to je da u prenaseljenim državama pravo na zdravstveno osiguranje ne može ostvariti većina stanovništva.

Reference: https://www.worldometers.info/