Predikcija stope samoubistava u pojedinim državama sveta

Vukašin Jović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
vukasin.jovic@uns.ac.rs

Milica Makarić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
makaric.milica@uns.ac.rs

Milan Lazić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
lazicy@uns.ac.rs

***Apstrakt*—Samoubistvo je složena pojava koja vekovima privlači pažnju raznih naučnika. Svake godine, samoubistvo je među 10 vodećih uzroka smrti u svetu među ljudima svih uzrasnih dobi. Kao ozbiljan zdravstveni problem zahteva našu pažnju, premda njegova kontrola i prevencija nisu nimalo jednostavna kako na njih mogu uticati razni faktori. Iz tog razloga, stalno postoje razne mere koje se preduzimaju u cilju pokušaja prevencije. Jedan od tih pokušaja sprovodi se korišćenjem računara, odnosno mašinskog učenja i raznih algoritama kako bi se izvršila predikcija i omogućilo sprečavanje samoubistava u svetu. Osnovni skup podakata koji sadrži osnovne socio-demografske podatke je proširen podacima koji se smatraju značajnim faktorima stope samoubistva. Ciljni skup podataka oformljen na taj način podeljen je na trening i test skup u proporciji 75%, 25%. Statističkom analizom je utvrđeno da je muški pol više sklon samoubistvu od ženskog pola, kao i da broj samoubistava opada u periodu od 1990. do 2016. godine.U sklopu ovog rada predloženi su različiti modeli regresije, kao i model klasifikacije. Za potrebe regresije su upotrebljeni sledeći modeli: Linear regression, Support Vector Regression, Gradient Boosted Tree, XGBoost i Random Forest, dok je kao primer klasifikacije upotrebljen Random Forest algoritam. Najbolji rezultat je postignut upotrebom regresije nad Random Forest modelom. Postignuta je R^2 mera od 0.97. Najveći uticaj na model imao je atribut koji predstavlja broj ljudi u populaciji države.**

Ključne reči—samoubistvo; predikcija; mašinsko učenje; regresija; stabla odlučivanja

I. UVOD

Od početka ljudske civilizacije ljudi su tragali za načinima da se suprotstave autodestruktivnom ponašanju i povećaju osećaj zadovoljstva životom. Samoubistvo se najčešće definiše kao autodestruktivni akt kojim individua prema sopstvenoj intenciji sama prouzrokuje smrt. [1] Ono kao izraz autoagresije u teorijskom i metodološkom proučavanju ne smatra se autohtonim fenomenom, već kategorijom heterogenih karakteristika. Ne može se posmatrati kao trenutni događaj već kao process koji traje. Samoubistvo ima specifičan zdravstveni i socijalni značaj jer ozbiljno učestvuje u kauzalitetu mortaliteta.

Samoubistvo je u skoro svim zemljama sveta među deset glavnih uzroka smrti. Na to upućuje podatak da oko million ljudi godišnje izvrši samoubistvo, dok između deset i dvadeset miliona pokuša samoubistvo. Putevi koji vode od agresije prema nasilju i samoubistvu nisu linearni, već su uslovljeni većim brojem interaktivnih kulturalnih, psiholoških, bioloških, socijalnih i situacionih faktora. [2] Neki od glavnih faktora su postojanje bolesti ili neke vrste poremećaja, prethodni pokušaji samoubistava, razne traume, okolina i slično međutim pristupanje tim podacima je vrlo teško i gotovo nemoguće, iako podaci o umrlim usled samoubistva postoje od XVIII veka. [3] Sa druge strane, postoje pristupačniji podaci koji su takođe među glavnim faktorima i korišćeni su u ovom radu a to su populacija zemlje, pol, starosno doba i prosečna primanja. Na osnovu znanja o ovim faktorima, postavlja se pitanje da li je moguće napraviti predikciju stepena samoubistva u različitim zemljama kako bi se što više moglo uticati na njihovo sprečavanje i pružanje pomoći što je više moguće.

Upravo ovim problemom se i bavi ovaj rad. U njemu će biti prikazano jedno rešenje za procenu predikcije samoubistva u određenim zemljama. Podaci na osnovu koga je obučen model predstavljaju praćenje broj samoubistava od 1990. do 2016. godine. Pored prethodno navedenih podataka koji su korišćeni u ovom radu, kao dodatni podatak upotrebljen je prosečan broj sunčanih dana. U brojnim studijama o suicide pronađena je povezanost suicida i sezonskih varijacija. Izražen je jasan uticaj gorišnjih doba na suicide pri čemu je najveći broj izvršen u toku proleća i leta, što potvrđuje da stopa suicida tokom vremena korespondira sa sezonskim varijacijama. [4]

Detaljniji opis podataka i rešenja izložen je u ostatku rada. Naredno poglavlje se bavi srodnim istraživanjima i sličnim problemima na ovu temu. Poglavlje 3 opisuje skup podataka korišćenog za obučavanje i validaciju modela. Pored toga objašnjen je postupak spajanja i obrađivanja sakupljenih podatka kako bi se dobio ciljni set podataka. Potom su u poglavnju 4 predstavljene metodologije i analize korišćene za rešavanje datog problema. Nakon toga sledi poglavlje 5 u kojem su prikazani primena algoritama i njihovi rezultati. Na kraju, poglavlje 6 sadrži izveden zaključak i diskusiju na temu rezultata.

II. PREGLED POSTOJEĆE RELEVANTNE LITERATURE

Prilikom istraživanja radi boljeg upoznavanja sa tematikom ovog rada, pronađeno je nekoliko radova u kojima je obrađen sličan problem. Neki od njih su navedeni u nastavku teksta.

U radu [5] je izvršena predikcija rizika pokušaja samoubistava kroz vreme korišćenjem različitih tehnika mašinskog učenja. Korišćene su tradicionalne i novije tehnike mašinskog učenja, kako bi se omogućilo poređenje rezultata. Kao predstavnik tradicionalnog načina korišćena je logistička regresija, a algoritam Random Forest je upotrebljen u svrhu novije tehnologije mašinkog učenja. Predikcija je vršena na osnovu medicinskih podataka iz kliničkih zdravstvenih kartona u kojima su zabeleženi slučajevi povreda za koje se zna ili sumnja da su samonanešene. Takođe, u obzir su uzeti i zdravstveni kartoni u kojima nisu zabeleženi podaci o pokušajima samoubistva. Skup podataka je sadržao podatke o 5167 pacijenata. Od parametara rad je uzeo u obzir demografske podatke poput godina, pola, nacinalnosti, ranijih dijagnoza, socio-ekonomskog statusa kao što su obrazovanje, imovina i podaci o zaposlenosti. Za evaluaciju modela korišćena je AUC (eng. *area under the receiver operating characteristic curve*) kao i preciznost (eng. *precision*) i odziv (eng. *recall*) metrike. Model je pokazao dobre rezultate sa sledećim rezultatima: AUC=0.84, precision=0.79, recall=0.95, i uspeo je da preciznost pogađanja svede sa 720 dana na 7 dana pre pokušaja samoubistva. Ovaj rad je pokazao uspešnost Random Forest algoritma za predikciju ovakvog tipa problema.

Rad [6] se bavi predikcijom ideja pojedinaca o samoubistvu na osnovu socio-demografskih, fizičkih i psiholoških obeležja. Predikcija se vrši upotrebom Random Forest algoritma. Od prvobitnih 47 obeležja, zapaženo je da je model obučen sa 39 karakteristika postigao najveću Kappa[1] vrednost. Potom je eliminacijom atributa unazad (eng. *backward selection*) odbrano 15 obeležja za koje se pokazalo da postižu neznatno nižu Kappa vrednost. Kao neki od najbitnijih parametara su izdvojeni: depresija, nivo stresa u svakodnevnom životu, pol, starost. Kako bi se izbeglo prekomerno uklapanje (eng. *overfitting*)[2] i povećala generalizacija modela, korišćena je 10-struka unakrsna validacija. Skup podataka od ukupno 11.628 pojedinaca (5.814 pojedinaca sa idejom o samoubistavu i isto toliko pojedinaca koji nisu imali ideje o samoubistvu) podeljen je na trening (10.466 pojedinaca) i test skup (1.162 pojedinaca) uz očuvanje odnosa 1:1 između dve klase. Predočeni model je postigao tačnost 0,81. Iz prethodno izloženog, uočava se pogodnost korišćenja Random Forest algoritma za rešavanje problema ovog tipa. Takođe, zaključuje se i da je značajno koristiti unakrsnu validaciju kako bi model bio uopšteniji.

U radu [7] je razvijen model za predviđanje pokušaja samoubistava južnokorejskih adolescenata na osnovu stabla odlučivanja. Skup podataka predstavlja kompoziciju socio-demografskih i psiholoških podataka ispitanika (stepen depresije, delikvencija, samopouzdanje, optimizam) i podatka da li je ispitanik u prethodnih godinu dana pokušao da izvrši samoubistvo. Stablo odlučivanja sa sobom povlači opasnost od potencijalnog overfitting-a u slučajevima kada je prisutna velika količina podataka. U ovom radu je bilo svega 2754 ispitanika, te je ta mogućnost mala. Takođe, veliki broj promenljivih može da utiče na dubinu stabla, što otežava klasifikaciju. Podaci u ovom radu su podeljeni u trening i test skup u odnosu 70/30. Validacija na osnovu testnog skupa je pokazala preciznost od 0.9. Kako je skup podataka za problem koji se opisuje u ovom radu mnogo veći od skupa podataka navedenog rada, nije preporučljivo da se koristi isti algoritam.

Rad [8] istražuje tezu da na suicidno ponašanje utiče sunčeva svetlost. Za istraživanje su korišćeni podaci o svim samoubistvima u Austriji u periodu od 1970-2010. godine. Podaci o prosečnom broju sunčanih satu u toku dana dobijeni su sa 86 reprezentativnih meteoroloških stanica. Kao rezultat istraživanja izvedena su dva zaključka. Prvi zaključak je .da postoji pozitivna korelacija između broja samoubistava i broja sunčanih sati na dan samoubistva, kao i do 10 dana pre samoubistva. Drugi zaključak je da postoji negativna korelacija između broja samoubistava i broja sunčanih sati tokom 14 do 60 dana pre samoubistva. Na osnovu ovih zaključaka, u skup podataka koji se obrađuje u ovom radu su dodati podaci o prosečnom broju sunčanih dana po godini za svaku državu.

III. OPIS SKUPA PODATAKA

Ovo poglavlje opisuje postupak nastanka finalnog seta podataka potrebnog za analizu i formiranje modela sistema. Inicijalni skup podataka predstavljao je skup preuzet sa veb stranice koja sadrži veliki broj različitih skupova podataka - kaggle.com. [9] U pitanju je skup podataka koji poredi socio-ekonomske informacije sa brojem samoubistava iz 101 zemlje u periodu od 1985. do 2016. godine. i on sadrži sledeće atribute (*u zagradi je prikazan tačan naziv labele*):

- naziv države (*country*)
- godina na koju se podaci odnose (*year*)
- starosna grupa preminulih (*age*)
- pol (*sex*)
- broj stanovnika države (*population*)
- bruto domaći proizvod države (*gdp_for_year ($)*)
- bruto domaći proizvod po glavi stanovnika države (*gdp_per_capita ($)*)
- broj samoubistava države za datu godinu, koji za potrebe ovog rada predstavlja ciljni atribut (*suicides_no*)
- broj samoubistava na 100 hiljada stanovnika države za datu godinu (*suicides/100k pop*)
- indeks humanog razvoja (*HDI for year*)
- naziv generacije (*generation*)
- naziv države sa godina (*country-year*)

Dalja analiza pomenutog skupa pokazala je da postoje nedostajuće vrednosti indeksa humanog razvoja. Iz tog razloga je ovaj podatak izbačen iz skupa podataka. Takođe su uklonjeni i podaci o nazivu generacije, nazivu države s godinom i bruto domaćem proizvodu po glavi stanovnika države. Naziv države sa godinom je uklonjen jer predstavlja redundantan podatak, naziv generacije iz razloga što nema doprinos u daljem

[1] Kappa vrednost je mera međusobnog slaganja obeležja

[2] Prekomerno uklapanje je greška u modelovanju koja se javlja kada je funkcija previše usko uklopljena na ograničeni skup podataka

Attributes	age	country	gdp_for_year ($)	population	salaries	sex	suicides/100k pop	suicides_no	sunshine_hours_per_year	year
age	1	-0.003	0.000	-0.164	0.000	0	0.051	-0.173	0.000	-0.003
country	-0.003	1	-0.037	0.097	-0.044	0	-0.047	-0.087	-1	0.125
gdp_for_year ($)	0.000	-0.037	1	0.848	0.035	0	-0.042	0.592	0.130	0.091
population	-0.164	0.097	0.848	1	0.012	0.012	-0.047	0.744	0.134	-0.001
salaries	0.000	-0.044	0.035	0.012	1	0	0.063	0.021	-0.127	-0.042
sex	0	0	0	0.012	0	1	-0.441	-0.191	0	0
suicides/100k pop	0.051	-0.047	-0.042	-0.047	0.063	-0.441	1	0.225	-0.170	-0.069
suicides_no	-0.173	-0.087	0.592	0.744	0.021	-0.191	0.225	1	0.065	-0.025
sunshine_hours_per_year	0.000	-1	0.130	0.134	-0.127	0	-0.170	0.065	1	-0.018
year	-0.003	0.125	0.091	-0.001	-0.042	0	-0.069	-0.025	-0.018	1

Slika 1. Matrica korelacije atributa

korišćenju skupa podataka kako predstavlja potpunu korelaciju sa starosnim grupama, dok je bruto domaći proizvod po glavi stanovnika uklonjen jer se ta vrednost može dobiti deljenjem bruto domaćeg proizvoda sa brojem stanovnika države. Pored toga, uklonjene su i države Azerbejdžan, Bosna i Hercegovina i Turska i svi njihovi podaci iz razloga što su postojale nedostajuće vrednosti u određenom broju godina.

Mnogi za pojam samoubistva, pored ostalih pomenutih faktora, vezuju i nizak životni standard odnosno mala primanja pojedinca. Iz tog razloga je analiziran skup podataka preuzet sa veb stranice Ekonomske komisije za Evropu (UNECE), koji sadrži informacije o prosečnim mesečnim primanjima država kroz godine. [10] U pitanju su prosečna mesečna primanja 56 država u periodu od 1990. do 2017. godine. U ovom i prethodnom skupu podataka jasna je razlika između broja država za koje postoje podaci te je prilikomspajanja skupova podataka, iz tog razloga, izbačeno 45 država za koje ne postoje adekvatni podaci. Pored toga, postoji razlika u opsegu godina te je urađen presek da bi se za finalni opseg posmatrao opseg od 1990. do 2016. godine.

Kako je pronađeno da je broj samoubistava najviše zabeležen u periodu proleće-leto te da klimatsko vreme značajno utiče na stepen samoubistva, u razmatranje je uzet i skup podataka o broju sunčanih sati po godini u gradovima sveta. Adekvatni skup podataka za države nije pronađen, te su podaci ručno preuzeti i obrađeni sa veb stranice wikipedia.org. [11] Dati podaci su bili grupisani po kontinentima i prikazani za svaki mesec u godini. Iz tog razloga se nakon preuzimanja vršilo sabiranje broja sunčanih sati po mesecu za gradove, čije države postoje u ciljnom skupu podataka i time je dobijem broj sunčanih sati svake države iz ciljnog skupa podataka.

Finalni skup podataka sadrži podatke o 56 država u periodu od 1990. do 2016. godine i njega čine sledeći atributi (u zagradi je prikazan tačan naziv labele):

- naziv države (*country*)
- godina na koju se podaci odnose (*year*)
- starosna grupa preminulih (*age*)
- pol (*sex*)
- broj stanovnika države (*population*)
- bruto domaći proizvod države (*gdp_for_year ($)*)
- broj samoubistava države za datu godinu, koji za potrebe ovog rada predstavlja ciljni atribut (*suicides_no*)
- broj samoubistava na 100 hiljada stanovnika države za datu godinu (*suicides/100k pop*)
- prosečna mesečna primanja (*salaries*)
- broj sunčanih sati (*sunshine_hours_per_year*)

Konačno, ciljni skup podataka koji se sastoji iz 11.885 redova je podeljen na trening i test podskupove u odnosu 75% - 25% i podela je izvršena tako da su u trening skupu podaci od 1990. do 2008. godine, a u test skupu od 2009. do 2016. godine. Ovako formiran trening skup podataka je upotrebljen za obučavanje modela kako bi se izvršila predikcija stope samoubistva. Pored toga, nad trening skupom korišćena je i unakrsna validacija koja služi za procenu modela algoritma na neviđenim podacima.

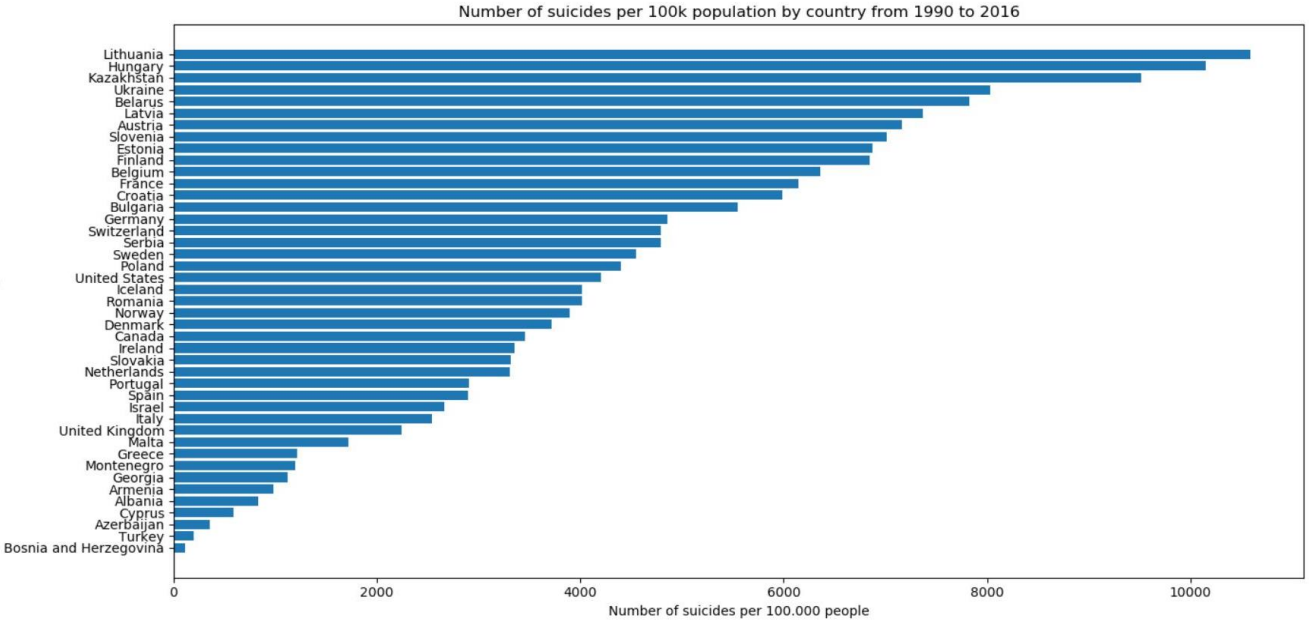
IV. STATISTIKA

U ovom poglavlju su opisane statističke analize koje su izvršene nad prethodno opisanom skupu podataka. Cilj ovih analiza je utvrđivanje postojanja veze između broja samoubistava i nekog drugog atributa iz celokupnog skupa podataka. Za utvrđivanje povezanosti između atributa upotrebljen je operator Correlation Matrix. Ovaj operator se koristi kako bi se izvršila analiza atributa odnosno njihova uzajamna povezanost (korelacija). Prikazan je u vidu matrice gde svaka ćelija predstavlja vrednost korelacije između dva atributa. U matrici se nalaze korelacije svakog para atributa koje su predstvaljene vrednostima od -1 do 1. Pozitivna vrednost korelacije označava da se vrednosti atributa proporcionalno povećavaju (pozitivna veza). Negativna vrednost korelacije predstavlja negativnu vezu, te se vrednosti parametara povećavaju obrnuto proporcionalno. Rezultat korišćenja ovog operanda je prikazan matrično (slika 1).

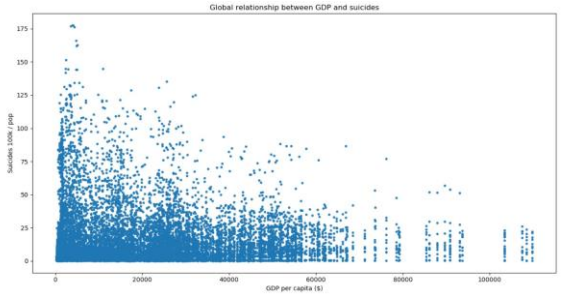
Najmanja vrednost korelacije atributa broj samoubistava (*suicides_no*) je uočena za atribut koji predstavlja prosečan iznos zarade (*salaries*). Kako je vrednost korelacije između ova dva atributa blizu nule, korelacija između njih praktično ne postoji.

Najveća vrednost korelacije je primećena između broja samoubistava (*suicides_no*) i broja stanovnika (*population*). To znači da što u državi ime više stanovnika, veći je i broj samoubistava. Sa druge strane, ukoliko se izvrši normalizacija broja samoubistava na 100.000 populacije zemlje dobiju se drugačiji rezultati. Upravo to je prikazano na slici 2 gde se vidi odnos broja samoubistava na 100.000 populacije zemalja iz ciljnog skupa podataka.

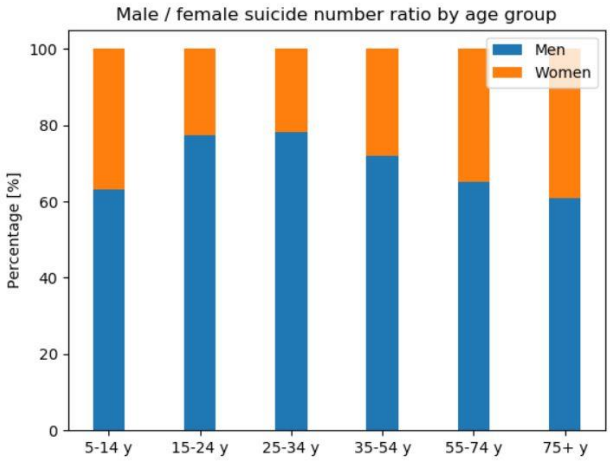
Daljim istraživanjem utvrđeno je da postoji veza između godišnjeg bruto domaćeg proizvoda i broja suicida na globalnom nivou, i to je prikazano na slici 3. U državama sa vrlo malim bruto domaćim proizvodom po broju stanovnika zabeležena je najveća stopa samoubistava i ta stopa opada sa njegovim porastom. Muškarci su više skloni samoubistvu od žena, u svim starosnim grupama. Takođe, broj samoubistava raste za oba pola do perioda od 35 do 54 godine, nakon čega počinje da opada što je i prikazano na slikama 4 i 5. Najznačajniji podatak jeste da globalni trend samoubistava opada od 1990. do 2015. godine i to je prikazano na slici 6.



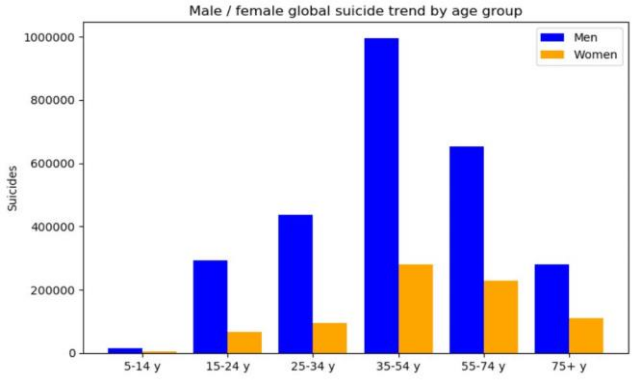
Slika 2. Globalni broj samoubistava na 100.000 populacije



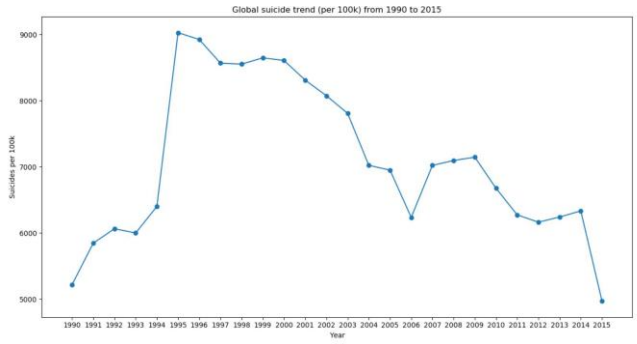
Slika 3. Veza bruto domaćeg prihoda i broja samoubistava



Slika 4. Procentualni odnos muškaraca i žena po starosnim grupama



Slika 5. Prikaz i rast broja samoubistava kod muškaraca i žena



Slika 6. Ukupan broj samoubistava na 100.000 populacije po godinama

V. METODOLOGIJA

Metodologija ovog rada se može podeliti u dve celine: priprema podataka i primena različitih modela za predikciju stope samoubistva.

Analizom je utvrđeno da je na prethodno objašnjenom skupu podataka bilo neophodno izvršiti neke izmene, kako bi podaci bili pripremljeni za obučavanje i testiranje modela. Kako su podaci za atribute country, sex i age prvobitno bili u tekstualnom obliku, nad njima je bilo potrebno primeniti transformaciju u brojeve. Prvobitno je upotrebljena funkcija nad trening i nad test skupom podataka čiji je rezultat izvršavanja dodavanje novih kolona za svaku državu, i postavljanje vrednosti 1 samo za državu iz tog reda. Za sve ostale države je postavljena vrednost 0. Tako dobijeni vektor se naziva *one-hot* vektor, tj. binarni vektor. Sa druge strane, korišćena je i druga funkcija koja je vrednosti svakog prethodno navedenog atributa transformisala u brojčane vrednosti. Konkretno, za vrednosti atributa country, naziv svake države je zamenjen indeksom te države u skupu podataka, te su tako nadalje države predstavljene brojevima od 0 do 43. Isto tako, atribut sex je nakon ove transformacije sadržao samo vrednosti 0 i 1, umesto vrednosti female i male. Atribut age je prezentovan vrednostima od 0 do 5, umesto sledećim vrednostima: 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years. Utvrđeno je da su svi isprobani modeli imali bolje performanse ukoliko su navedeni atributi transformisani na drugi po redu opisani način.

Sledeća transofrmacija koja je izvršena jeste normalizacija. Ovaj postupak je izvršen jer određeni algoritmi, kao što je na primer Support Vector Regression, imaju slabije performanse ukoliko podaci nisu normalizovani. Normalizacija je primenjena na sve podatke, kako bi se proverilo da li ima uticaj na rezultate predikcije.Normalizacija je izvršena pomoću formule:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

gde z predstavlja normalizovanu vrednost koja je nalazi u opsegu [0, 1], dok je x atribut koji se normalizuje.

Međutim, u primeru koji je obrađen u ovom radu, pokazalo se da su predikcije svih primenjenih algoritama lošije nakon što je primenjena normalizacija. Stoga su podaci vraćeni u oblik u kom su bili pre ove transformacije.

Nad podacima je primenjena i PCA (eng. *Principal component analysis* – Analiza glavnih komponenti) procedura, koja omogućava smanjenje velikog broja atributa. PCA koristi ortogonalnu transformaciju da pretvori skup obeležja eventualno koreliranih promenljivih u skup vrednosti linearno nekoreliranih promenljivih koje se nazivaju glavnim komponentama. Kako u primeru ovog rada skup podataka sadrži ukupno 8 obeležja na osnovu kojih je vršena predikcija, PCA algoritam nije u velikoj meri smanjio broj obeležja. U skladu sa tom konstatacijom, broj komponenti (*n_components_*) koje su nastale kao rezultat rada PCA iznosi 7.

Za predikciju stope samoubistava odabrano je nekoliko različitih regresionih modela, kao i jedan klasifikacioni model. Regresioni modeli koji su korišćeni su:

- Linearna regresija
- Linearna regresija sa Ridge regularizacijom
- Linearna regresija sa Lasso regularizacijom
- Support Vector Regression
- Gradient Boosted Tree
- XGBoost
- Random Forest

U svrhu klasifikacionog modela je upotrebljen ansambl klasifikacionih stabala u Random Forest konfiguraciji.

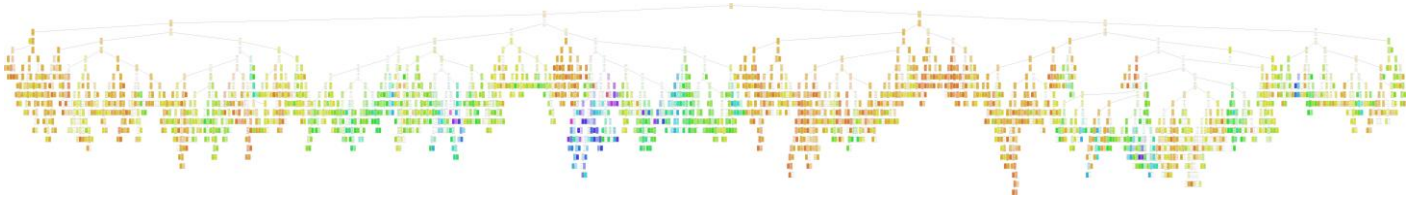
Validacija za svaki od navedenih regresionih modela je uređena pomoću unakrsne validacije (eng. *cross validation*). Za potrebe unakrsne validacije u ovom radu, korišćena je *K-fold* validacija, te je skup podataka za trening podeljen na 5 podskupova. Unakrsna validacija je primenjena nad trening skupom podataka, pre primene modela na test skup podataka. Kao mera evaluacije performansi regresionih modela korišćena je R2

mera kao i koren srednje vrednosti kvadrata greške (eng. *root_mean_squared_error*).

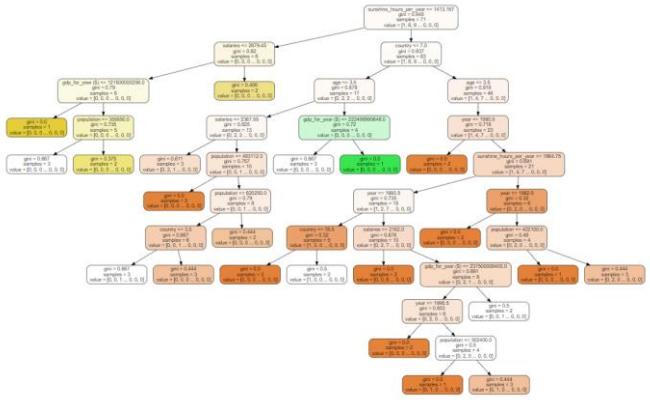
Prvi korak je bio izbor parametra alpha za Ridge[3] i Lasso[4] regularizaciju. Regularizacija predstavlja tehniku koja se koriste za smanjenje greške pravilnim postavljanjem funkcije na zadati trening set i izbegavanje overfitting-a. Međutim, nakon izvršene optimizacije, nije bilo značajnog poboljšanja performansi u odnosu na linearnu regresiju bez regularizacije. Rezultati predikcije su se razlikovali za svega 0.0001. Na osnovu tog zaključka, regresioni modeli sa Ridge i Lasso regularizacijom su isključeni iz daljeg razmatranja.

Model Support Vector Regression je optimizovan po pitanju parametra C koji se odnosi na regularizaciju. Optimalna vrednost ovog parametra za opisani skup podataka je 1000.

Gradient Boosted Tree model je optimizovan po pitanju parametara: broj estimatora (*n_estimators*), random_state, stopa učenja (*learning_rate*) i dubina stabla (*max_depth*). Optimalne vrednosti ovih parametara su: n_estimators= 20, random_state = 42, learning_rate = 0.1 i max_depth = 4.



Slika 7. Prikaz celog stabla Random Forest modela



Slika 8. Prikaz dela stabla Random Forest modela

Model XGBoost je optimizovan po pitanju parametara: broj estimatora (*n_estimators*), stopa učenja (*learning_rate*) i dubina stabla (*max_depth*). Vrednosti koje su pronađene kao optimalne su: broj estimatora 50, stopa učenja 0.1 i dubina stabla 4.

Random Forest model je takođe optimizovan. Optimizacija je urađena za sledeće parametre: broj estimatora (*n_estimators*), random_state i dubina stabla (*max_depth*). Optimizovane vrednosti ovih parametara iznose respektivno: 50, 42 i 100. Na slici 7 je prikazano celokupno stablo Random Forest modela koji je korišćen u ovom radu, dok je na slici 8 prikazan samo deo stabla kako bi se bolje uočio proces treniranja modela.

Optimizacija parametara za sve regresione modele je izvršena pomoću GridSearchCV funkcije.

Da bi se regresioni model mogao prebaciti u klasifikacioni, neophodno je bilo izvršiti odgovarajuće izmene nad skupom podataka. Prvobitna zamisao je bila da se atribut koji predstavlja broj samoubistava (*suicides_no*) podeli na određeni broj klasa, te bi na taj način približno isti broj samoubistava pripadao jednoj klasi. Međutim, u posmatranom atributu je jako velika distribucija vrednosti, te bi u tom slučaju postojao jako veliki broj klasa. Da bi se to izbeglo, za ciljnu labelu je izabran atribut koji predstavlja skaliranu vrednost broja samoubistava na 100.000 ljudi (*suicides/100k pop*). Klase su formirane tako da prvoj klasi pripadaju vrednosti ciljne labele od 0 do 10, drugoj od 11 do 20, i tako redom. Za evaluaciju klasifikacionog modela korišćena je F1 mera (eng. *F1 score*), kao i parametri preciznost (eng. *precision*) i odziv (eng. *recall*). Preciznost predstavlja udeo dobro predviđenih instanci neke klase, u ukupnom broju instanci koje je model svrstao u datu klasu. Drugim rečima, pokazuje koliki deo rezultata u jednoj klasi je uspešno klasifikovan. Odziv pokazuje osetljivost modela, odnosno pokazuje koliko je relevantnih rezultata algoritam vratio. Ovaj broj pokazuje udeo dobro predviđenih instanci neke klase, u ukupnom broju instanci koje zapravo pripadaju toj klasi. F1 mera kombinuje preciznost i odziv, i računa se po sledećoj formuli:

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

[3] Ridge je tehnika za analizu višestrukih regresijskih podataka koji pate od multikolinearnosti koristeći najmanje procene kvadrata za najveću tačnost između procenje i tačne vrednosti

[4] Lasso je metoda regresijske analize koja vrši selektivni izbor i regularizaciju u cilju povećanja tačnosti predviđanja i interpretacije statističkog modela koji proizvodi

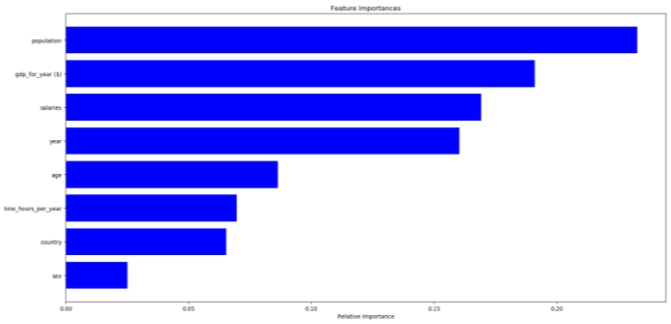
VI. REZULTATI

Rergresioni modeli, kao i klasifikacioni model, testirani su na test skupu podataka koji je izdvojen iz finalnog skupa podataka na način koji je objašnjen u poglavlju 3. Evaluacija je izvršena koristeći optimalne parametre dobijene u procesu optimizacije modela koji je opisan u poglavlju 5. Rezultati evaluacije prikazani su u tabeli 1.

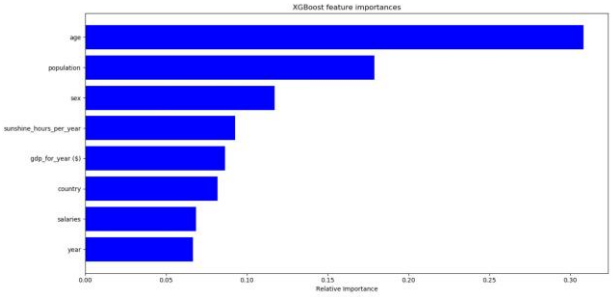
Problem regresije	R^2	root mean squared error
Linearna regresija	0.65	493.40
Support Vector Regression	0.26	704.37
Gradient Boosted Tree	0.10	780.22
XGBoost	0.40	668.38
Random Forest	0.97	125.54

Tabela 1**.** Rezultati regresije na test skupu

Na osnovu rezultata prikazanih u tabeli 1, zaključuje se da je za problem opisan u ovom radu najbolje performanse postigao Random Forest algoritam. R2 mera za ovaj algoritam iznosi 0.97, što je veoma blizu maksimalnoj vrednosti. Što se tiče i koren srednje vrednosti kvadrata greške, vrednosti koje su prikazane u tabeli 1 su na prvi pogled relativno velike. Međutim, ciljna labela ovog problema broj samoubistava koja sadrži veoma velike vrednosti (maksimalna vrednost na nivou celog skupa podataka ovog atributa je 11767). Oslanjajući se na to, koren srednje vrednosti kvadrata greške svih modela je prihvatljiv. Na slici 9 je prikazano koji procenat važnosti ima svaki atribut kod Random Forest modela, dok je na slici 10 prikazana istovetna analiza ali pri korišćenju XGBoost algoritma.



Slika 9. Procenat važnosti atributa kod Random Forest algoritma



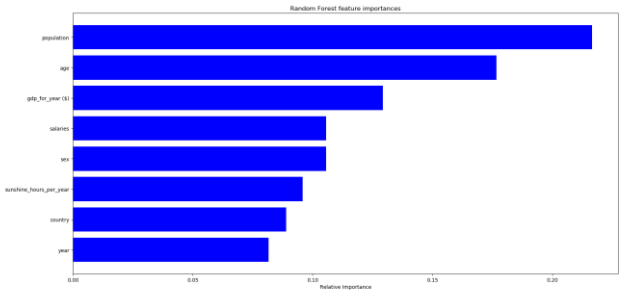
Slika 10. Procenat važnosti atributa kod XGBoost algoritma

Za problem klasifikacije je, kao što je već rečeno u poglavlju 5, kao mera evaluacije korišćena F1 mera, preciznost i odziv. Rezultati koji su ostvareni nad test podacima prilikom klasifikovanja prikazani su u tabeli 2.

Problem klasifikacije	Random Forest
F1 score	0.75
precision	0.75
recall	0.75

Tabela 2. Rezultati regresije na test skupu

Kao što se vidi u tabeli 2, F1 mera, preciznost i odziv imaju identičnu vrednost 0.75. Obzirom da maksimalna vrednost parametra F1 iznosi 1, dobijeni rezultati su veoma dobri. Na slici 11 se vidi kojim redom i sa kojim procentom su atributi bili značajni prilikom primene klasifikacionog modela Random Forest.



Slika 11. Procenat važnosti atributa kod klasifikacije

Posmatrajući zajedno i regresioni i klasifikacioni problem, bolji rezultat je ostvarila regresija. Kao razlog za slabije rezultate klasifikacije može se navesti to što je kao ciljna labela korišćen atribut broj samoubistava na 100.000 populacije (*suicides/100k pop*), umesto neskaliranog atributa broj samoubistava (*suicides_no*). Pored toga, treba imati u vidu i matricu korelacije prikazanu na slici 1, gde se vidi da atribut suicides_no ima veću koreliranost sa ostalim podacima od atributa suicides/100k pop.

VII. ZAKLJUČAK

Problem kojim se bavi ovaj rad jeste predviđanje stope samoubistava u određenim državama sveta gde su se, pored osnovnih podataka o tim državama, koristili i neki specifični faktori poput prosečnih mesečnih primanja građana i broja sunčanih sati u godini. Rešenje ovog problema dovelo bi do znanja koji to faktori dovode individuu do situacije da počini samoubistvo što bi moglo pomoći u prevenciji takvih situacija.

Prvi korak ka tom rešenju bio je formiranje odgovarajućeg skupa podataka nad kojim su kasnije sprovedene dalje analize. Nad formiranim skupom prvo je izvršena statistička analiza kako bi se videla korelacija obeležja i istakle moguće zavisnosti. Pored toga, izvršen je grafički prikaz trenda pojedinih faktora na broj samoubistava čime je potvrđen veći deo postavljenih hipoteza o uticaju određenih faktora na stepen samoubistava. Potom je putem regresije i klasifikacije obučeno više modela gde je za regresiju kao mera evaluacije korišćeno odstupanje srednje-kvadratne devijacije(eng. *Root mean square error*, RMSE) i R2, dok je za klasifikaciju korišćena f1 mera. Utvrđeno je da je najbolji model Random Forest, koji je postigao vrednost R2 od 0.97.

Dalji razvoj rešenja obuhvata dobavljanje podataka o državama, dobavljanje podataka o godinama pre 1990. i nakon 2016. godine kao i korišćenje drugih algoritama za predikciju. Takođe, skup podataka se može proširiti sledećim značajnim faktorima:

- zdravstveni problemi pojedinaca
- bračni status
- stanje države
- najzastupljenija vera u državi
- stepen obrazovanja
- nezaposlenost
- mesto stanovanja
- rastureni dom
- kriminalni nagoni
- sukobi u porodici
- stepen usamljenost
- stepen anksioznosti
- delikvencija
- depresija
- stepen satisfakcije
- broj prethodnih pokušaja samoubistava
- agresija
- sezonske promene
- klimatske promene i dr.

VIII. LITERATURA

[1] Dragana Ljušić, „Uticaj zdravstvenog stanja na promenu ponašanja u periodu pre samoubistva“, 2017.

[2] Priručnik, „Prevencija samoubistva i samoubilačkog ponašanja mladih”, Institut za mentalno zdravlje, Beograd 2010.

[3] Chesnais JC, „Les morts violentes dans le monde”, Population & Sociétés 2003; (395): 1-4.

[4] Milić Č, „Sezonske varijacije - faktor rizika za nastanak suicida”, Medicinski pregled Novi Sad 2010.

[5] Colin G. Walsh, Jessica D. Ribeiro, Joseph C. Franklin: „Predicting Risk of Suicide Attempts Over Time Through Machine Learning”, 2017.

[6] Seunghyong Ryu, Hyeongrae Lee, Dong-Kyun Lee, and Kyeongwoo Park: „Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population“, 2018.

[7] Sung Man Bae, Seung A Lee, Seung-Hwan Lee: „Prediction by data mining, of suicide attempts in Korean adolescents: a national study“, 2015.

[8] Benjamin Vyssoki, Nestor D. Kapusta, Nicole Praschak-Riede, Georg Dorffner, Matthaeus Willeit: „Direct Effect of Sunshine on Suicide“, 2014.

[9] „Kaggle: Your Machine Learning and Data Science Community” [Online] Available: https://www.kaggle.com/. [Accessed: 25-Apr-2020].

[10] „Gross Average Monthly Wages by Country and Year”, [Online] Available: https://w3.unece.org/. [Accessed: 27-Apr-2020].

[11] „List of cities by sunshine duration“, [Online] Available: https://en.wikipedia.org/wiki/List_of_cities_by_sunshine_duration. [Accessed:30-Apr-2020]