Analiza i predikcija stope samoubistava po državama

Marko Katić E2 06/2019
Računarstvo i automatika
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
katicmarko96@gmail.com

Aleksandar Stanić E2 07/2019
Računarstvo i automatika
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
sale96@protonmail.com

Boris Bibić E2 26/2019
Računarstvo i automatika
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
borisbibic1996@gmail.com

Apstrakt*—Ljudski život nema cenu. Uprkos tome, svakih četrdeset sekundi ugasi se jedan život u svetu zbog samoubistva. Stope samoubistava su u nekim zemljama još uvek visoke ili pokazuju trend rasta, ipak uočava se pad vrednosti stopa samoubistava u mnogim državama sveta. Ideja priloženog rada jeste prepoznavanje najznačajnijih faktora rizika samoubistava i kreiranje modela za predikciju stopa samoubistava. Ove modele bi mogle da koriste države u cilju razvoja strategija i nacionalnih planova radi smanjenja trenutnih stopa smrtnosti ili da zaustave trenutne trendove rasta istih. Kako bi se realizovala navedena ideja, prikupljeni su skupovi podataka iz brojnih istraživanja dostupnih na internetu za period od 1990. do 2017. godine. Izvršen je proces analize i obrade podataka kako bi isti bili validni za predikciju. Na ovaj način su dobijene različite verzije skupova podataka koje su korišćene za treniranje više različitih verzija modela. Zbog velike dimenzionalnosti podataka vršeno je ispitivanje redukcije podataka na tačnost prediktivnih modela. Za redukciju podataka su korišćeni algoritmi Low Variance Filter, Random Forest, Factor Analysis i Principal Component Analysis, a za predikciju Elastic Net, XGBoost i Partial Least Squares. Za predikciju samoubistava, na osnovu korišćenih podataka, najbolje se pokazao Partial Least Squares. Rezultati o najznačajnijim faktorima rizika su dobijeni uz pomoć Random Forest algoritma i poklapaju se sa rezultatima medicinskih istraživanja.

Ključne reči*—samoubistava; rudarenje podataka; redukcioni algoritmi; prediktivni algoritmi.

1. UVOD

Svake godine u svetu blizu 800000 ljudi izvrši samoubistvo, a stručnjaci ukazuju da je na svako samoubistvo najmanje 20 pokušaja samoubistva. Prema statistikama Svetske zdravstvene organizacije samoubistvo je drugi najčešći uzrok smrti kod osoba uzrasta od 15 do 29 godina [1]. Prevencija samoubistava se zasniva na prepoznavanju visoko rizičnih grupa, te je obaveza i odgovornost svake države da radi na njihovom otkrivanju i zbrinjavanju. Prepoznavanje rizičnih faktora i njihov uticaj na određenu populaciju je od krucijalne važnosti za prevenciju jednog od vodećih uzroka smrti mladih. Ovaj projekat bi pomogao zakonodavcima da bolje razumeju faktore rizika za samoubistva u njihovim državama, kako bi preuzeli adekvatne mere prevencije. Takođe, ovaj projekat bi bio od značaja i stručnjacima koji se bave prevencijom i prepoznavanjem samoubistava kod pojedinaca u riziku na osnovu prepoznatih faktora rizika u njihovim zemljama.

Cilj rada je kreiranje modela za predikciju stope samoubistava po državama na osnovu obeležja koja predstavljaju najznačajnije faktore rizika, a koja su dobijena nakon redukcije skupova podataka. Skupovi podataka su prikupljeni na osnovu stručne literature i po izboru autora. Prikupljeni skupovi podataka su analizirani i obrađeni radi spajanja u jedan sveobuhvatni skup podataka koji se dalje koristio za treniranje prediktivnih modela. Nedostajuće vrednosti sveobuhvatnog skupa su popunjavane *Elastic Net* algoritmom [2] i uz pomoć dubokog učenja (engl. *Deep Learning*) [3]. Zbog velike dimenzionalnosti podataka redukcija je vršena sledećim algoritmima: *Low Variance Filter* [4], *Random Forest* [5], *Factor Analysis* [6] *i Principal Component Analysis* [7]. Sa redukovanim i sveobuhvatnim skupovima podataka trenirani su sledeći prediktivni modeli: *Elastic Net, XGBoost* [8] *i Partial Least Squares* [9]. Dobijeni rezultati pokazuju da je najveći koeficijent determinacije (R^2) prediktivnih modela 93,6% dok je istovremeno srednja kvadratna greška (engl. *Mean Square Error - MSE*) 0,049. Rezultati o najznačajnijim faktorima rizika su dobijeni uz pomoć *Random Forest* algoritma i poklapaju se sa rezultatima medicinskih istraživanja.

U poglavlju dva su predstavljena slična rešenja koja su se bavila ispitivanjem faktora rizika i kreiranjem prediktivnih modela za samoubistva. Poglavlje tri se bavi procesom izrade željenog rešenja kao i opisom skupova podataka, metoda i algoritama koji su korišćeni. Četvrto poglavlje predstavlja analizu dobijenih rezultata sa diskusijom. Poglavlje pet

predstavlja sumarizaciju čitavog rada i diskusiju o mogućnostima unapređenja.

2. PRETHODNA REŠENJA

Među pronađenim rešenjima izdvaja se rad „*Risk factors for suicide ideation among adolescents: five-year national data analysis.*“ [10] koji se bavi identifikacijom faktora rizika za nastajanje suicidnih misli kod adolescenata u Južnoj Koreji. Analizirano je 370568 studentskih odgovora na pitanja vezana za samoubistva u periodu od 5 godina. Pored upitnika analizirani su i demografski podaci (pol, ekonomski status, život bez jednog ili oba roditelja,…), faktori rizika za poremećaj mentalnog zdravlja (depresija, manjak sna, visok stres, konzumiranje alkohola...). Koristio se algoritam „*cross-sectional descriptive design*“ sa (sekundarnom) statističkom analizom, bez upotreba metoda mašinskog učenja. Studija je pokazala da su adolescenti bez jednog ili oba roditelja, kao i oni u nižem socioekonomskom statusu u većem riziku od samoubistva. Takođe, studija je potvrdila značajnost mentalnog zdravlja za prevenciju samoubistva. Konzumiranje alkohola, pušenje i seksualne aktivnosti su povezane sa suicidalnim mislima, po ovoj studiji.

Autori Choi, Soo Beom i drugi su u svojoj studiji [11] istraživali faktore rizika i napravili prediktivni model za pokušaje samoubistava kod južnokorejske populacije. Analizirano je 1567 muškaraca i 3726 žena starijih od 20 godina koji su imali suicidalne misli u okviru nacionalne ankete koja je trajala od 2007. do 2012. godine. Od ovih ispitanika 106 muškaraca i 188 žena je pokušalo izvršiti samoubistvo. Korišćen je algoritam „Multi variate logistic regression analysis with backward stepwise elimination”. Kao faktori rizika za pokušaj samoubistva kod muškaraca su odabrani starost, obrazovanje, rak i depresivni poremećaj. Starost, obrazovanje, osnovna sredstva za život, ograničenje svakodnevnih aktivnosti, depresivni poremećaj, stres, pušenje i redovna fizička aktivnost su odabrani kod žena. Tačnost modela predviđanja iznosi 0,728 kod muškaraca, a kod žena 0,716. Za razliku od autora [11], u ovom radu je početni broj izabranih faktora veći, dok su najznačajnije faktore odabrali sami algoritmi.

Cilj studije [12] je da istraži uticaj različitih osobina samopovređivanja koje nisu dovele do samoubistva (engl. *non-suicidal self-injury - NSSI*) na suicidalne ideje (engl. *suicidal ideas - SI*), planiranje suicida (engl. *suicide planning - SP*) i pokušaj samoubistva (engl. *suicide attempt - SA*). Analizirano je 359 studenata sa istorijom *NSSI*, gde su posmatrani demografski podaci, depresija i 58 *NSSI* karakteristika, kao i istorija *SI*, *SP* i *SA*. Korišćeni su algoritmi *Elastic Net* regresija, *Decision tree* i *Random Forest.* Za predviđanje *SI* i *SP* najveći uticaj imaju simptomi depresije i anti-suicidalne funkcije (zaustavljanje ili izbegavanje sucidalnih misli). Na *SA* najviše su uticale anti-suicidalne funkcije, *NSSI* vezani medicinski tretmani i samopovređivanje. U studiji [12] su faktori rizika bili vezani za medicinsku anamnezu, dok se u radu autora akcenat stavlja na globalne faktore kao što su nivo korupcije i inflacija u državi. Medicinska anamneza nije toliko dostupna za istraživanje, analizu i kreiranje prediktivnih modela kao što su dostupni podaci Ujedinjenih Nacija, što je i mana rada [12].

Autori rada [11] su uradili sličnu studiju pod nazivom „*Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea.*“ [13]. Ova studija je želela da ispita verovatnoću uspešnog samoubistva koristeći osnovne karakteristike osoba i istoriju posete medicinskoj ustanovi. Analizirano je 819951 subjekata iz Nacionalnog zavoda za zdravstveno osiguranje Južne Koreje. Podaci su podeljeni na trening i validacioni skup. Korišćeni su algoritmi *Cox regresija* (engl. *Cox regression*), metod potpornih vektora (engl. *Support Vector Machine - SVM*) i duboke neuronske mreže (engl. *Deep Neural Networks – DNN)* za kreiranje prediktivnih modela. Od ukupnog broja ispitanika, 2546 ljudi je umrlo od namernih samopovreda tokom praćenja. Faktori rizika: pol, starost, vrsta osiguranja, prihod domaćinstva, invalidnost i medicinska evidencija (uključujući mentalne poremećaje i poremećaje ponašanja) izabrani su *Cox regresijom* sa postupnim eliminacijama unazad kao bitna obeležja. Tačnost *Cox regresije* je 68,8%, *SVM* 68,7%, a *DNN* 68,3%.

Razlika između rada autora i gore navedenih radova jeste i u tome da pronađeni radovi obuhvataju samo jednu državu, dok su autori rada obuhvatili veći broj država.

3. METODOLOGIJA

Metodologija ovog rada uključuje tri oblasti: rad sa skupom podataka, redukciju dimenzionalnosti skupa podataka i treniranje modela za predikciju. U daljem tekstu biće detaljno opisan svaki korak koji je uključen u metodologiju ovog rada.

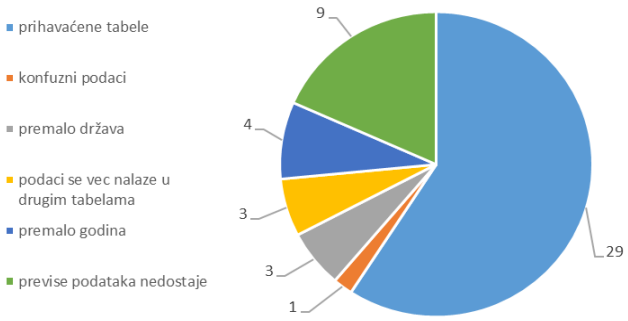
3.1 Prikupljanje skupova podataka

Polazni skup podatka sadrži broj samoubistava po državama od 1990. do 2017. godine sa 6468 podataka na osnovu kojih je treniran model. Ostali skupovi podataka obuhvataju podatke iz oblasti državnog uređenja, ekonomije, prava, medija, kao i drugih relevantnih oblasti koje mogu predstavljati potencijalne faktore rizika samoubistva. Većina skupova podataka je odabrana na osnovu istraživanja stručne literature, dok je ostatak izabran kao potencijalno interesantan za istraživanje od strane autora.

3.2 Analiza i priprema podataka

Inicijalni skup podataka je prikupljen sa interneta, gde je najveći broj podataka preuzet sa stranica worldbank.org [14] i ourworldindata.org [15]. Ovaj skup čini 49 *Excel* i *CSV* fajlova. Za dalji rad bilo je neophodno da svaki skup podataka sadrži *ISO3* kôd države, godinu i obeležja koja su karakteristična za taj skup podataka.

Nakon pripreme podataka, sprovedena je eksplorativna analiza. Alatom *RapidMiner* [16] uočeno je da najveći broj skupova podataka ima najmanje nedostajućih obeležja u rasponu od 1990. do 2017. godine. Tabele, tj. skupovi podataka koji su imali malo obeležja u ovom opsegu su izbačeni iz daljeg razmatranja. Takođe, tabele sa malo država, tabele sa konfuznim obeležjima (vrednosti nisu logične za dato obeležje) ili tabele koje sadrže obeležja koja se nalaze u drugim tabelama su izbačene iz dalje obrade. U dalju obradu je uključeno 29 tabela, dok je 20 tabela odbačeno, a detaljnija statistika oko odbacivanja tabela se može videti na slici 1.



Slika 1. Raspodela primenljivosti tabela nakon obrade

Većina od preostalih 29 skupova podataka je imala sledeće probleme: nedostajuća obeležja za pojedine godine i države, nedostajuće godine i/ili države, duplikate godina i/ili država, nedostajuće vrednosti su bile prikazane nulom ili je opseg godina bio izvan 1990.-2017. Duplikati godina su uočeni na histogramima dobijenim *RapidMiner* alatom i potom su uklonjeni istim alatom. Kod skupova gde su nulom predstavljane nedostajuće vrednosti, one su bile obrisane i time postale zaista nedostajuće vrednosti. Skupovi podataka koji su imali nedostajuće vrednosti su propuštani kroz *Python* skriptu koja je vršila popunjavanje nedostajućih vrednosti na četiri različita načina:

- *Fill with 0* - Sve nedostajuće vrednosti zamenjene su sa nulom.
- *Fill with min* - Sve nedostajuće vrednosti zamenjene su sa najmanjom vrednosti iz tog skupa obeležja.
- *Fill with max* - Sve nedostajuće vrednosti zamenjene su sa najvećom vrednosti iz tog skupa obeležja.
- *Fill with mean* - Sve nedostajuće vrednosti zamenjene su sa srednjom vrednosti iz tog skupa obeležja.

Ukoliko skup podataka nije imao neophodan opseg godina bilo je potrebno napraviti prediktivni model za vrednosti nedostajućih godina. Za ovu potrebu korišćeni su *XGBoost* i *Elastic Net* regresioni algoritmi. Pripremljeno je osam različitih prediktivnih modela koji su nastali treniranjem regresionih algoritama sa skupovima podataka čije su nedostajuće vrednosti popunjene kombinacijom gorenavedenih metoda. Za svaki skup podataka izabran je najbolji prediktivni model. Bilo je neophodno pronaći i nedostajuće godine za svaku državu, zbog čega je kreiran rečnik čiji je ključ bila oznaka države, a vrednost lista nedostajućih godina. Prediktivni model je predviđao vrednosti obeležja za svaki nedostajući element u rečniku. Pored *XGBoost* i *Elastic Net* regresije, urađena je i regresija sa *Deep Learning* modelom u *RapidMiner* alatu. Ovim su dobijene dve vrste skupova podataka u zavisnosti od prediktivnog algoritma: *Elastic Net* i *Neural Net* verzija. Skupovi koji su imali godine van opsega, zbog povećanja preciznosti prediktivnih modela, su nakon predikcije filtrirani da sadrže samo godine u opsegu od interesa. Skupovi podataka kojima nije bila potrebna regresija su, takođe, filtrirani na isti način.

3.3 Spajanje skupova podataka

Nakon obrade skupova, usledilo je njihovo spajanje uz pomoć *RapidMiner* alata. Ukupan broj država nakon spajanja je bio 231. Uočeno je prilikom spajanja da određeni broj država, koje se ne nalaze u svim tabelama, kao vrednost obeležja iz tih tabela imaju prazno polje. Procenjeno je da države koje imaju nedostajuća obeležja treba izbaciti, pošto bi popunjavanje tih obeležja unelo previše nepouzdanih podataka. Nakon izbacivanja, broj država u *Elastic Net* verziji skupa podataka je iznosio 142. Tokom popunjavanja nedostajućih vrednosti sa *Deep Learning* algoritmom, uočeno je da pojedine države imaju premalo vrednosti koje bi mogle biti iskorišćene za treniranje, pa su te države izbačene. Time je broj država u *Neural Net* verziji spojenog skupa podataka iznosio 123.

3.4 Obeležja spojenog skupa podataka

Nakon spajanja skup podataka sadrži 33 obeležja:

- *Coutry Code* - *ISO3* kôd svake od država u skupu
- *Year* - godina na koju se podatak odnosi
- *Agricultural methane emissions* - emisija metana u poljoprivredi
- *corruption_index* - procenat korupcije u državi
- *Daily caloric supply* - dnevni unos kalorija po glavi stanovnika
- *Death rate* - godišnja stopa smrtnosti na 1000 stanovnika
- *Access to electricity* - procenat populacije koji ima struju
- *Employers, total (% of total employment)* - procenat ljudi koji su vlasnici privatnog biznisa od ukupnog broja zaposlenih u državi.
- *Total score (Press)* - ocena slobode medija data od strane organizacije *Freedom House* u godišnjem *Freedom of the Press* izveštaju.
- *Human Rights Protection Scores* - ocena poštovanja ljudskih prava
- *inflation* - godišnja inflacija
- *Happines* - ocena sreće stanovnika date države
- *Military expenditure* - finansijska sredstva koju država troši na oružane snage
- *Political Regime* - ocena političkih režima
- *Population density* - gustina naseljenosti po kvadratnom kilometru
- *PTS (Political terror scale)* - prisustvo političkog terora u državi. Postoji 5 nivoa gde je prvi najbolji, a peti najgori.
- *Alcohol and substance use disorders* - procenat stanovništva koji ima problem sa alkoholom i drogom

- *Total origin* - procena ukupnog broja emigranata po državi porekla
- *Unemployment* - udeo radno sposobnog stanovništva koji nije zaposlen
- *Urban population growth* - rast gradskog stanovništva
- *Dominant religion* - dominantna religija u datoj državi
- *Region* - region u kojem se nalazi data država
- *IncomeGroup* – pripadnost grupi u odnosu na ukupna primanja stanovništva
- *Per capita CO_2 emissions* - emisija ugljen dioksida po glavi stanovnika
- *Fertility rate* - broj dece koji žena rodi tokom svog života
- *Mediascore* - ocena slobode medija u državi
- *Internet usage* - procenat stanovništva kojem je dostupan internet
- *Eating disorders* - procenat populacije sa poremećajima u ishrani
- *Depressive disorders* - procenat stanovništva koji boluje od depresije
- *Mental and substance use disorders* - procenat stanovništva koji ima mentalne poremećaje i poremećaje zavisnosti
- *Deaths - Self-harm* - procenat stanovništva koji umre od samopovređivanja (ovo je obeležje koje se prediktuje)

3.5 Enkodovanje string obeležja

Prediktivni algoritmi za vrednosti svakog obeležja zahtevaju numeričku vrednost. Pošto u spojenim skupovima podataka postoje obeležja sa string vrednostima (npr. obeležje *Country Code*) potrebno je izvršiti enkodovanje podataka. Dve vrste enkodovanja podataka su razmatrana za ovaj rad:

- *Label Encoding* - Enkoduje labele sa vrednostima između 0 i *n*-1, gde *n* predstavlja broj različitih labela. Ukoliko se labela ponovi u istoj koloni, ona dobija istu vrednost koja joj je dodeljena ranije.
- *One-Hot Encoding* - Enkoduje tako što svaku kolonu sa *n* kategoričkih podataka razdvaja na *n* kolona. Vrednosti su zamenjene sa jedinicama i nulama u zavisnosti od toga koji red ima tu vrednost.

Problem *Label Encoding*-a jeste da prilikom treniranja, prediktivni model može zaključiti da enkodovana kolona ima linearnu zavisnost sa kolonom koja se prediktuje, što nije tačna pretpostavka. Tako model može doći do pogrešnog zaključka, na primer što je veći broj u koloni *Country Code*, onda je i veća stopa samoubistva. Stoga je za enkodovanje podataka korišćen *One-Hot Encoding*. Ovo je značajno podiglo dimenzionalnost spojenih skupova podataka. Skupu podataka dobijenog *Elastic Net* algoritmom je nakon enkodovanja povećana dimenzionalnost na 188 obeležja, dok je skupu podataka dobijenog sa dubokom neuronskom mrežom dimenzionalnost povećana na 169 obeležja.

3.6 Redukcija dimenzionalnosti

Za redukciju dimenzionalnosti autori su odlučili da koriste algoritme:

- *Low Variance Filter* - Za svaku kolonu se gleda varijacija promene njenog sadržaja i ukoliko je varijacija manja od neke granice (engl. *threshold*), smatraće se da loše utiče na dalji rad. Primer dobre primene ovog algoritma jeste izbacivanje kolona sa konstantnim vrednostima. Sprovedene su redukcije spojenih skupova podataka sa četiri različite granice procenta varijacije promene: 20%, 40%, 60% i 80%.
- *Random Forest* - Radi na osnovu *stabla odlučivanja*, tako što prilikom kreiranja „slučajne šume“, uzima jedan podskup obeležja da istrenira novo *stablo odlučivanja* i to ponavlja dok ne napravi „šumu“ od njih. Na osnovu generisanih *stabala odlučivanja*, algoritam izračunava koja obeležja su najbitnija. *Random Forest* se koristio za analizu najznačajnijih faktora rizika samoubistava.
- *Principal Component Analysis (PCA)* - Jedan je od najpoznatijih algoritama za redukciju dimenzionalnosti. Algoritam iz skupa podataka pronalazi glavne komponente (nova obeležja izvedena iz ulaznih ortogonalnom transformacijom) i njih vraća kao rezultat. Prilikom redukcije skupova podataka, podešava se parametar koji predstavlja broj glavnih komponenti na koji algoritam treba da svede originalni skup podataka. Za skup podataka dobijen *Elastic Net* regresijom korišćene su četiri različite vrednosti parametra *PCA* algoritma: 47, 94, 141 i 188, odnosno 42, 84, 126 i 169 za skup podataka dobijen sa dubokom neuronskom mrežom. Ovo predstavlja redukciju broja obeležja originalnog skupa za 75%, 50%, 25% i 0%.
- *Factor Analysis* - Uz *PCA*, *Factor Analysis* je takođe jedan od najčešće korišćenih algoritama za redukciju dimenzionalnosti. Ova tehnika je efikasna kada u skupu podataka postoji puno visoko koreliranih obeležja, jer algoritam grupiše takva obeležja i pravi *faktore* (nova obeležja) za svaku grupaciju. Slično *PCA*, prilikom redukcije podešava se broj *faktora* na koji će *Factor Analysis* redukovati ulazni skup podataka. Za ovaj rad, korišćene su iste vrednosti kao kod redukcije sa *PCA* algoritmom.
- *Partial Least Squares* (*PLS*) - Prediktivni algoritam koji prilikom kreiranja vrši redukciju dimenzionalnosti sličnu *PCA* metodi, osim što *PLS* uzima u obzir i izlazne vrednosti skupa podataka prilikom redukcije. Prilikom implementacije ovog algoritma prosleđivan je celi skup podataka i svi redukovani skupovi podataka koji su dobijeni prethodno spomenutim algoritmima redukcije dimenzionalnosti.

Pored izabranih algoritama za redukciju dimenzionalnosti, razmatrani su i sledeći algoritmi: *High Correlation Filter*,

Backward Feature Elimination, Foward Feature Construciton, Autoencoder, Linear Discriminant Analysis, t-Distributed Stochastic Neighbor Embedding (*t-SNE*), *Idepedent Component Analysis, Uniform Manifold Approximation and Projection* (*UMAP*), *Lasso Regularisation, Ant Colony Optimization, Particle Swarm Optimization, Genetic Algorithm* i *Sequential Floating Selection*. Razlozi zbog kojih ovi algoritmi nisu odabrani su različiti: nemogućnost pronalaženja *Python* implementacije, nepouzdanost redukcije (*Autoencoder*), vremenska efektivnost, nemogućnost dobrog prikaza rezultata redukcije i gubitak informacija.

3.7 Treniranje prediktivnih modela

Korišćeni algoritmi za predikciju stope samoubistava su:

- *XGBoost* - Algoritam se odlikuje dobrim performansama, brzinom i mogućnosti paralelnog rada i omogućava rad sa većim skupovima podataka.
- *Elastic Net* - Predstavlja kombinaciju *Rigid* i *Lasso* regresije. Pogodan je za situacije sa velikim brojem parametara gde nije poznato koliko određeni parametar utiče na rezultat.
- *Partial Least Squares* (*PLS*) – Algoritam koji pored redukcije dimenzionalnosti vrši i treniranje prediktivnog modela sa najboljim redukovanim skupom podataka. *PLS* je dao najbolje rezultate u radu autora.

Pored navedenih prediktivnih algoritama razmatran je i *Cox regression* algoritam. Isti je izbačen zbog nemogućnosti rada sa skupom podataka velike dimenzionalnosti.

Zbog validacije prediktivnih modela svi skupovi podataka su podeljeni na trening i test skup. Trening skup se koristio za obučavanje prediktivnih modela i sadrži 80% podataka iz skupa. Preostalih 20% čini test skup za validaciju dobijenih modela. Treba napomenuti da skupovi podataka koji se koriste prilikom treniranja prediktivnih modela ne smeju imati nedostajuće (*null*) vrednosti.

3.8 Validacija modela

Po završenom treniranju prediktivnih modela urađena je validacija modela sa test skupom podataka. Za evaluaciju performansi modela korišćene su dve metode:

- *Srednja kvadratna greška* (engl. *Mean Square Error - MSE*) - Služi za dobijanje prosečne kvadratne razlike između procenjenih i stvarnih vrednosti. Drugim rečima, *MSE* služi kao mera kvaliteta prediktivnih modela. Ova mera kvaliteta uvek daje nenegativne vrednosti, gde su vrednosti bliže nuli bolje.
- *Koeficijent determinacije* (R^2) - Koristi se kod modela čija je glavna svrha predviđanje budućih ishoda na osnovu srodnih informacija. R^2 daje meru koliko dobro model predviđa. Vrednosti su najčešće u rasponu od 0 do 1, a što su bliže jedinici to je model bolji u predviđanju. Koeficijent determinacije može imati negativnu vrednost, što pokazuje da model loše predviđa buduće ishode.
- *Unakrsna validacija (engl. cross-validation*) – Ova metoda izračunava sposobnost modela da predviđa u zavisnosti od broja komponenti koje se nalaze u modelu. *Cross-validation* je korišćen kod *PLS* algoritma.

3.9 Optimizacija parametara

Parametri algoritama za redukciju dimenzionalnosti optimizovani su empirijski, ali je uzeta u obzir i domenska osnova. Za *Low Variance Filter* domenska preporuka je 80%, a u obzir su još uzeti 20%, 40% i 60%. Kod *PCA* i *Factor Analysis* algoritma, optimizacija broja željenih obeležja na koji se svodi dimenzionalnost je bila isključivo empirijska.

Elastic Net ima parametar *alpha*, kojim se podešava kompleksnost modela (od *Ridge* do *Lasso* regresije) kako ne bi došlo do *overfitovanja* modela. Domenska preporuka je da taj parametar bude 0,01, ali isprobane su i vrednosti 0,001, kao i 0,1 i 0,8.

XGBoost omogućava podešavanje parametara *learning_rate, colsample_bytree, max_depth, n_estimators*. *Learnig_rate* određuje veličinu koraka pri svakoj iteraciji dok se kreće ka minimalnoj funkciji gubitka. Optimalne vrednosti za *XGBoost learning_rate* su između 0,01 i 0,1, u radu je korišćena vrednost 0,1. *Colsample_bytree* određuje broj kolona koje koristi svako stablo, kako neka kolona ne bi previše uticala na rezultat. Preporučene vrednosti su između 0,3 i 0,8 ako ima puno kolona i ako je korišćeno *One-Hot* enkodovanje. U radu parametar *colsample_bytree* je postavljen na 0,3. *Max_depth* predstavlja dubinu svakog stabla, odnosno maksimalni broj obeležja koji se koriste u svakom stablu. Kreće se od vrednosti 3, koja se inkrementuje sve dok više nema poboljšanja u performansama, u radu je korišćena vrednost 5. *N_estimators* je broj *Gradient boosted* stabala. Podrazumevana i u radu korišćena vrednost je 10.

PLS algoritam na osnovu uzlaznog skupa podataka obavlja redukciju dimenzionalnosti, treniranje modela i validaciju rezultata sa *MSE*, R^2 i unakrsnom validacijom.

4. REZULTATI I DISKUSIJA

Rezultat rada jeste prepoznavanje najznačajnijih faktora rizika visoke stope samoubistava i kreiranje prediktivnog modela koji predviđa stopu samoubistva po državama. Ciljno obeležje koje se prediktuje je „*Deaths - Self-harm"* i predstavlja stopu samoubistava.

Svaki prediktivni model (*XGBoost, Elastic Net* i *PLS*) istreniran je sa pet skupova podataka: četiri skupa podataka dobijena od četiri algoritma redukcije dimenzionalnosti i jedan neredukovan skup podataka.

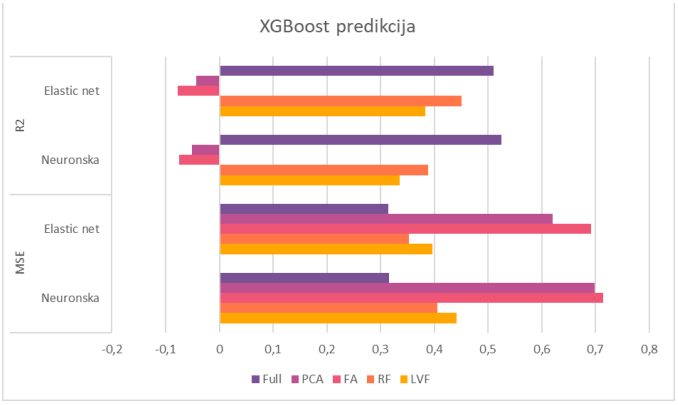
4.1 Rezultati prediktivnih modela

Rezultati R^2 i *MSE* validacije tri prediktivna modela su prikazani na slikama 2, 3, i 4. Uočava se da je najlošije rezultate imao *XGBoost*, dok je najbolje imao *PLS*. Poredeći dve verzije skupa podataka, *Elastic Net* verzija je imala bolje rezultate kod *XGBoost* i *Elastic Net* regresije, dok je *Neural Net* verzija dala bolje rezultate kod *PLS* regresije. Razlike između ove dve verzije skupova podataka su u proseku ispod 4%.

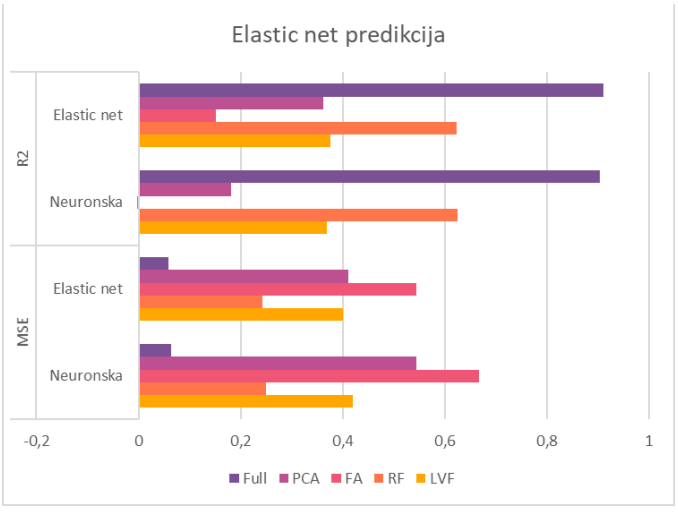
Najniži *MSE* je kod *PLS* modela za *PCA* i neredukovanu verziju skupa podataka i iznosi 0,049. Najviši R^2 je 93,6%, takođe kod *PLS* modela za *PCA* verziju skupa podataka. *PLS* model treniran sa neredukovanom verzijom skupa podataka je sledeći sa najvišim R^2 koji iznosi 93,5%.

Neredukovana verzija je uglavnom imala najbolje rezultate, osim u slučajevima *PLS* modela, gde su *Factor Analysis* i *PCA* imali bolje R^2 rezultate za 0,1%, što pokazuje da je redukcija dimenzionalnosti u skoro svim slučajevima uticala negativno na tačnost modela. Kada se posmatra koeficijent determinacije, kod *PLS* prediktivnog modela, razlika između najbolje i druge najbolje verzije skupa podataka je minimalna, dok se kod druga dva prediktivna modela ta razlika kreće od 5,89% do 28,62%.

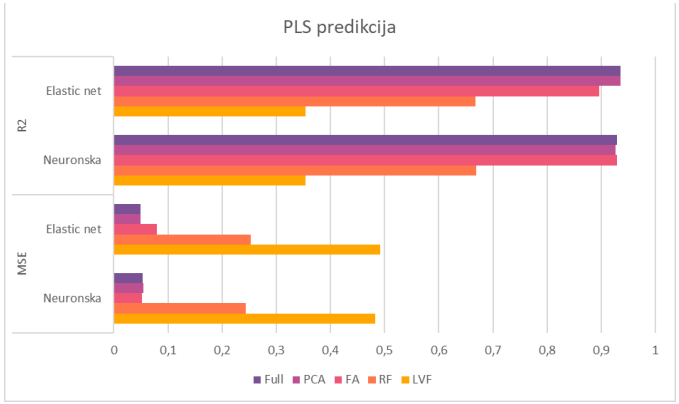
Kada se posmatraju verzije skupova podataka na osnovu redukcionog algoritma, prosečan koeficijent determinacije je sledeći: 30,37% za *Factor Analysis*, 36,19% za *Low Variance Filter*, 38,51% za *PCA*, 57,08% za *Random Forest* i 78,55% za neredukovani skup podataka. Ovo još jednom potvrđuje da se sa neredukovanim skupom podataka dobija najveća preciznost. *Random Forest* je pokazao najveću sposobnost redukcije dimenzionalnosti skupova podataka nezavisno od ulaznog skupa i prediktivnog algoritma jer ima najbolju prosečnu vrednost koeficijenta determinacije (57,08%).



Slika 2. Validacioni rezultati za *XGBoost* model

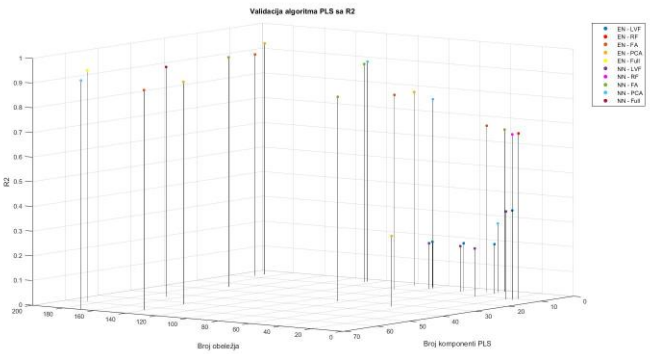


Slika 3. Validacioni rezultati za *Elastic Net* model

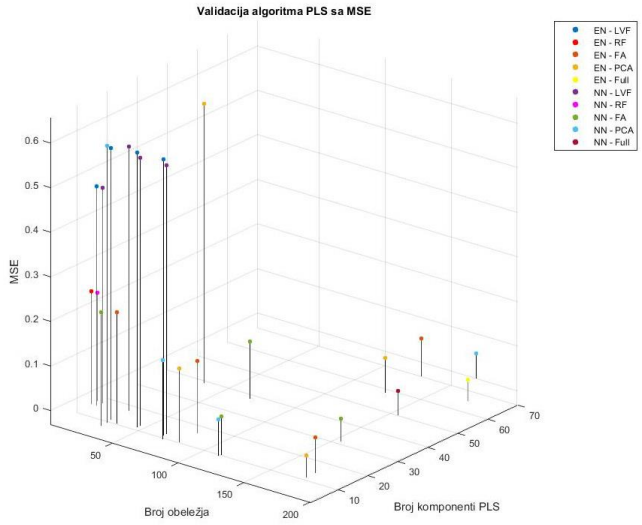


Slika 4. Validacioni rezultati za *PLS* model

Cross-validation je korišćen za validaciju *PLS* modela. Na slikama 5 i 6 prikazan je najbolji *cross-validation* rezultat za svaki od redukcionih modela i verzija skupova podataka. Broj obeležja ulaznog skupa podataka je predstavljen na *x*-osi, broj komponenti koji je *PLS* algoritam označio kao optimalan se nalazi na *y*-osi. Na *z*-osi su predstavljeni R^2 (slika 5) i *MSE* (slika 6). Primećuje se da se najbolji rezultati postižu kada se *PLS* algoritmu proslede neredukovani skupovi podataka i prepusti algoritmu da izvrši redukciju i pronađe optimalni broj komponenti. U većini slučajeva taj broj komponenti je manji od 20, što pokazuje da je to uglavnom dovoljan broj komponenti za predikciju samoubistava. Na osnovu unakrsne validacije najbolje rezultate su dali *Factor Analysis* i *PCA* redukcioni algoritmi, kao i neredukovani skup podataka. Nisu primećene značajne razlike između *Elastic Net* i *Neural Net* verzija skupova podataka.



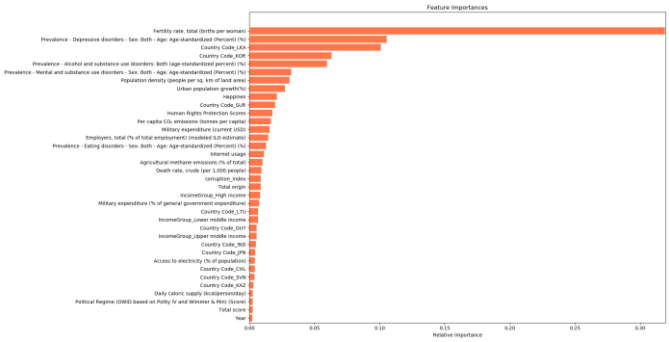
Slika 5. Unakrsna validacija *PLS* algoritma sa R^2



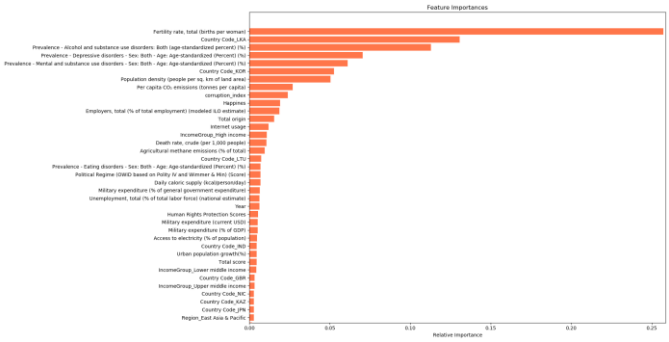
Slika 6. Unakrsna validacija *PLS* algoritma sa *MSE*

4.2 Najznačajniji faktori rizika samoubistava

Pored redukcije dimenzionalnosti, *Random Forest* vraća informaciju o najznačajnijim obeležjima. Na slikama 7 i 8 prikazan je deo grafikona koje je *Random Forest* algoritam generisao, a koji prikazuju značaj obeležja za *Elastic Net* i *Neural Net* verziju skupova podataka.

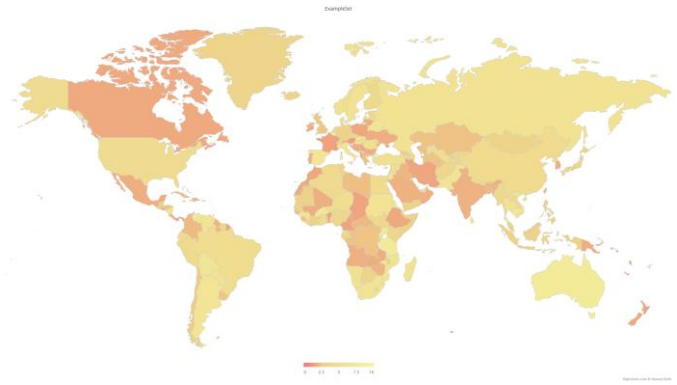


Slika 7. Graf značaja obeležja za *Random Forest* algoritam u *Elastic Net* verziji skupa podataka

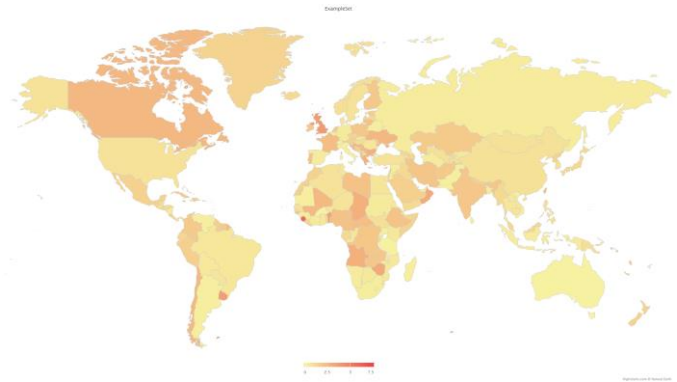


Slika 8. Graf značaja obeležja za *Random Forest* algoritam u *Neural Net* verziji skupa podataka

Kod *Elastic Net* skupa podataka najznačajnije obeležje je *Fertility rate*, potom redom slede *Depresive disoreders, life in Sri Lanka, life in South Korea* i *Alcohol and Substance abuse.* Kod *Neural Net* skupa podataka najznačajnije obeležje je *Fertility rate*, potom redom slede *life in Sri Lanka, Alcohol and Substance abuse, Depresive disoreders* i *Mental and substance use disorders.* Uočeno je da je *Fertility rate* u obe verzije skupa podataka najznačajniji faktor rizika samoubistava. Slika 9 prikazuje mapu sveta gde države označene crvenom bojom imaju mali *Fertility rate*, dok one označene sa žutom bojom imaju veći. Slika 10 prikazuje stopu samoubistava, gde su države sa visokom stopom prikazane crvenom bojom, a one sa niskom prikazane žutom. Ovo pokazuje da postoji obrnuta korelacija između broja dece koje žena rodi tokom svog života i broja samoubistava.



Slika 9. Prikaz *Fertility rate*-a na mapi sveta



Slika 10. Prikaz stope samoubistva na mapi sveta

4.3 Diskusija

Oko 60% skupova podataka prikupljenih sa interneta ispunilo je kriterijume i ušlo u uzorak. Najčešći razlog odbacivanja skupova podataka su bili nedostajući podaci (18,37%). Od ukupnog broja prihvaćenih tabela kod 62,1% njih je bilo potrebno popuniti nedostajuće podatke. Ovo svakako može imati uticaja na tačnost prediktivnih modela.

U poređenju sa rezultatima prethodnih rešenja [11] i [13], autori rada su koristili druge prediktivne algoritme i dobili bolje rezultate (93,6% u odnosu na 72,8% i 68,8%). Korišćenje većeg skupa podataka daje mogućnost dobijanja preciznije predikcije.

PCA je pokazao najbolje rezultate od ispitanih redukcionih algoritama kada je predikcija vršena sa *PLS* algoritmom, dok je *Random Forest* pokazao najveću sposobnost redukcije dimenzionalnosti skupova podataka nezavisno od algoritma predikcije.

U autorskom radu su dobijeni modeli čija je tačnost preko 90% kao što su *Factor Analysis* treniran sa *Neural Net* verzijom i *PCA* treniran sa *Elastic Net* verzijom.

Najznačajniji faktori rizika samoubistva koji su prikazani na slikama 7 i 8 se poklapaju sa faktorima koji su prepoznati u stručnoj literaturi [17]. Ipak neke od država sa visokim stopama samoubistava nisu prepoznate u *Random Forest* algoritmu, kao što je Grenland.

5. ZAKLJUČAK

Samoubistvo je visoko zastupljen javnozdravstveni problem, ne samo u svetu već i u Srbiji. Kreiranje modela za predikciju stope samoubistava i utvrđivanje najznačajnijih faktora rizika samoubistava može pomoći zakonodavcima i stručnjacima da prepoznaju i bolje razumeju faktore rizika za samoubistva u pojedinim državama. Utvrđivanje najznačajnijih faktora rizika samoubistava može pomoći državama u preduzimaju adekvatnih mera prevencije, kao i da identifikuju vulnerabilne grupe i pojedince u riziku. Sve države u svetu imaju obavezu da donesu strateške i akcione planove za smanjenje stopa smrtnosti samoubistava.

Problemu se pristupilo prvo istraživanjem potencijalnih faktora rizika samoubistava. Pronađeno je 49 različitih skupova podataka. Radi daljeg spajanja podataka izvršeno je refaktorisanje skupova podataka. Eksplorativnom analizom je izdvojen optimalan opseg godina i tabela koje mogu ići u dalju obradu. Dopunjeni su nedostajući podaci u skupovima podataka tako da podaci obuhvataju izabran opseg godina (1990-2017). Dopuna je izvršena *Elastic Net* regresijom i sa dubokim neuronskim mrežama, čime su dobijene dve verzije skupa podataka. Nakon dopune nedostajućih podataka, skupovi podataka su spojeni u jedan veliki skup podataka. Izbačene su države koje su posle spajanja imale nedostajuća obeležja. Izvršeno je enkodovanje string obeležja sa One-Hot enkodovanjem. Kao rezultat ovoga dimenzionalnost skupova podataka se povećala sa 33 obeležja na 188 za *Elastic Net*, odnosno 169 za *Neural Net* verziju. Zbog povećanja dimenzionalnosti skupova podataka urađena je redukcija. Odabrani su sledeći algoritmi za redukciju: *Low Variance Filter*, *Principal Component Analysis*, *Random Forest* i *Factor Analysis*. Pomoću ovih algoritama dobijeni su redukovani skupovi podataka koji su bili spremni za fazu treniranja prediktivnih modela. Od prediktivnih algoritama korišćeni su *XGBoost* regresija, *Elastic Net* regresija i *Partial Least Squares*. Svaki prediktivni model se trenirao sa četiri redukovana skupa podataka, kao i sa neredukovanim skupom podataka.

Validacija rezultata se vršila uz pomoć *MSE*, R^2 i unakrsne validacije. Najbolji rezultati su dobijeni kada se *PLS* prediktivni model trenirao sa *PCA* redukovanim skupom podataka (koeficijent korelacije je 93,6% i *MSE* 0,049). Neredukovana verzija je postigla najbolje rezultate osim kada je *PLS* prediktivni algoritam treniran sa *Factor Analysis* i *PCA* redukovanim verzijama skupova podataka. Ovo pokazuje uticaj redukcije na tačnost predikcije. *PLS* unakrsna validacija je najčešće redukovala prosleđeni neredukovani skup podataka na manje od 20 komponenti. Nisu primećene značajne razlike u tačnosti prediktivnih modela treniranih sa *Neural Net* i *Elastic Net* verzijama skupova podataka. Pomoću *Random Forest* algoritma generisan je graf sa merom značajnosti svakog obeležja. Među najznačajnijim faktorima rizika samoubistava su *Fertility rate, Alcohol and Substance abuse, Depresive disoreders* i *Mental and substance use disorders.*

Rezultati koji se odnose na najznačajnije faktore rizika, dobijenih sa *Random Forest* algoritmom, se većinom slažu sa podacima koji su dostupni u stručnoj literaturi. Države koje danas imaju visoke stope samoubistava, kao što su Južna Koreja, Šri Lanka i Litvanija, i u predikcijama modela imaju visoke stope samoubistava.

Pored ove činjenice, prednost rada jeste i velik skup podataka koji uključuje 188 država i vremenski period od 28 godina (1990.-2017. godina). Rad ima upotrebnu vrednost i stručnjaci ga mogu iskoristiti za dalja istraživanja kretanja stopa samoubistava na nacionalnom nivou, kao i za dalji istraživački rad u cilju unapređenja samog modela.

Potencijalni nedostaci rada su skupovi podataka koji su uključeni u uzorak, međutim sam kvalitet i celokupnost istih nije bila zadovoljavajuća. Primećeno je da se pojedine države sa visokom stopom samoubistava nisu našle visoko na listi *Random Forest* algoritma.

Postoji nekoliko tačaka mogućeg unapređenja rada: dodatne verzije redukcionih i prediktivnih algoritama, dodatni algoritmi za popunjavanje nedostajućih vrednosti, kao i uvođenje novih obeležja. Rad bi mogao da se unapredi u cilju regionalnog ili kontinentalnog praćenja i analize faktora rizika i uticaja na kretanje stopa samoubistava.

REFERENCE

[1] World Health Organization – Mental health: https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/ [Online]

[2] Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), pp.301-320.

[3] Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural networks*, *61*, pp.85-117.

[4] Cox, K.A., Dante, H.M. and Maher, R.J., Philip Morris USA Inc, 1993. Product appearance inspection methods and apparatus employing low variance filter. U.S. Patent 5,237,621.

[5] Ho, T.K., 1995, August. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

[6] Harman, H.H., 1976. Modern factor analysis. University of Chicago press.

[7] Jolliffe, I.T., 2002. Principal components in regression analysis. Principal Component Analysis, pp.167-198.

[8] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[9] Höskuldsson, A., 1988. PLS regression methods. Journal of chemometrics, 2(3), pp.211-228.

[10] Im, Y., Oh, W.O. and Suk, M., 2017. Risk factors for suicide ideation among adolescents: five-year national data analysis. *Archives of psychiatric nursing*, *31*(3), pp.282-286.

[11] Choi, S.B., Lee, W., Yoon, J.H., Won, J.U. and Kim, D.W., 2017. Risk factors of suicide attempt among people with suicidal ideation in South Korea: a cross-sectional study. BMC public health, 17(1), p.579.

[12] Burke, T.A., Jacobucci, R., Ammerman, B.A., Piccirillo, M., McCloskey, M.S., Heimberg, R.G. and Alloy, L.B., 2018. Identifying the relative importance of non-suicidal self-injury features in classifying suicidal ideation, plans, and behavior using exploratory data mining. Psychiatry research, 262, pp.175-183.

[13] Choi, S.B., Lee, W., Yoon, J.H., Won, J.U. and Kim, D.W., 2018. Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. Journal of affective disorders, 231, pp.8-14.

[14] The World Bank Group: https://www.worldbank.org/ [Online]

[15] Our World in Data: https://ourworldindata.org/ [Online]

[16] RapidMiner: https://rapidminer.com/ [Online]

[17] World Health Organization - Suicide: https://www.who.int/news-room/fact-sheets/detail/suicide [Online]