Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/16BE21Bibic

ANALIZA I PREDIKCIJA STOPE SAMOUBISTAVA UPOTREBOM TEHNIKA ISTRAŽIVANJA PODATAKA

SUICIDE RATE ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES

Boris Bibić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Priložen rad opisuje istraživanje o prepoznavanju najznačajnijih faktora rizika samoubistava i kreiranja modela za predikciju stopa samoubistava. Prikupljeni su skupovi podataka nad kojima je izvršen proces analize, obrade i spajanja podataka. Dobijene različite verzije skupova podataka su korišćene za treniranje više različitih verzija modela. Zbog velike dimenzionalnosti podataka vršeno je ispitivanje redukcije podataka na tačnost prediktivnih modela. Pronađeni su faktori rizika na globalnom i državnom nivou, a ispitan je i uticaj geografskih regija, religija i primanja na faktore rizika. Prediktivni modeli su evaluirani, a rezultati najznačajnijih faktora rizika su upoređeni sa rezultatima medicinskih istraživanja.*

Ključne reči: *Predikcija samoubistava, rudarenje podataka, predikcioni algoritmi, algoritmi za redukciju dimenzionalnosti, najznačajniji faktori rizika samoubistava*

Abstract – *This paper describes research on identifying the most significant risk factors for suicide and creating a model for predicting suicide rates. Data sets were collected over which the process of data analysis, processing and merging was performed. The obtained different versions of the data sets were used to train several different versions of the model. Due to the large dimensionality of the data, the reduction of data to the accuracy of predictive models was examined. Risk factors at the global and national levels were found, and the influence of geographical regions, religions and income on risk factors was examined. Predictive models were evaluated, and the results of the most significant risk factors were compared with the results of medical research.*

Keywords: *Suicide prediction, data mining, predictive algorithms, dimensionality reduction algorithms, the most significant suicide risk factors*

1. UVOD

Samoubistvo je ozbiljan javnozdravstveni problem koji disproporcionalno pogađa sve države sveta. Prema statistici Svetske zdravstvene organizacije (SZO) suicid je treći najčešći uzrok smrti kod dece uzrasta 15-19 godina, a drugi najčešći uzrok smrti kod osoba uzrasta 15-29 godina [1]. Prepoznavanje faktora rizika koji doprinose pojavi suicida je od krucijalne važnosti kako za zakonodavce tako i za stručnjake koji se bave prevencijom i prepoznavanjem samoubistva kod pojedinaca. Ovaj rad bi pomogao pri prepoznavanju faktora rizika u pojedinačnim državama i omogućio kreiranje procena kretanja stopa samoubistva.

Ideja istraživanja jeste prepoznavanje najvažnijih faktora rizika i kreiranje modela za predikciju samoubistva u svetu. Faktori rizika koji su razmatrani u radu su odabrani uz pomoć stručne literature i sličnih radova, ali neki su dodati od strane autora kao potencijalno interesantni. Prikupljeno je 49 tabela koje su analizirane, a izvučena statistika i informacije su pomogle pri daljoj obradi tabele i pronalaženju optimalnog opsega godina. Opseg od 1990. do 2017. godine je uzet kao optimalni za dalje razmatranje. Tabele sa premalo podataka u željenom opsegu su izbačene iz dalje obrade, pa je broj tabela spao na 29. Dalja obrada je obuhvatala popunjavanje nedostajućih vrednosti, gde su isprobani regresioni algoritmi *Elastic net* i *XGBoost*, kao i duboke neuronske mreže (engl. *DNN - Deep Neural Networks*). Za potrebe treniranja ovih modela, nedostajući podaci u skupovima podataka su popunjeni sa najmanjom i najvećom vrednosti iz skupa, sa nula vrednosti i sa prosečnom vrednosti iz skupa, čime su dobijene četiri varijacije skupova podataka. Nakon treniranja regresionih modela, originalne tabele su propušteni kroz *Elastic net* ili *XGBoost* model, u zavisnosti od toga koji je imao veću tačnost za posmatranu tabelu. Takođe, tabele su propuštene i kroz *DNN* model, pa su time dobijene dve varijacije početnih tabela bez nedostajućih vrednosti. Ove tabele su filtrirane da sadrže samo godine iz odabranog opsega (1990.-2017.).

Dobijene tabele su spojene po zajedničkim atributima, a to su godina i *ISO 3* kôd države, očuvajući varijacije tabela. Dobijeni su *Classic* i *Neural* verzija konačnog skupa podataka koje su se koristile za treniranje predikcionih modela za predviđanje stopa samoubistava, ali i za pronalaženje najznačajnijih faktora rizika. Broj atributa u obe verzije konačnog skupa je 33, a broj država je 142 za *Classic* i 123 za *Neural* verziju konačnog skupa podataka.

Usled velikog broja atributa u konačnom skupu podataka, rađena je redukcija dimenzionalnosti. Ispitana su 4 algoritma za redukciju dimenzionalnosti - *Low Variance Filter* (*LVF*), *Random Forest* (*RF*), *Principal Component Analysis* (*PCA*) i *Factor Analysis* (*FA*). Obe verzije konačnog skupa podataka su propuštene kroz ova četiri

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, vanr. prof.

algoritma čime je dobijeno osam novih (redukovanih) verzija konačnog skupa podataka. Ovih osam verzija, zajedno sa originalne dve verzije konačnog skupa podataka, su iskorišćene za treniranje tri prediktivna algoritma. *Elastic net, XGBoost* i *Partial Least Squares* (*PLS*) su odabrani za kreiranje prediktivnih modela.

Validacija prediktivnih modela je izvršena uz pomoć srednje kvadratne greške i koeficijenta determinacije, dok je za *PLS* dodatno rađena unakrsna validacija. *PLS* regresioni model treniran sa *PCA* redukovanim *Classic* skupom podataka se pokazao kao najbolji prediktivni model sa koeficijentom determinacije 93,6%. Unakrsna validacija je pokazala da je *PLS* algoritam davao najbolje rezultate kada je radio sa neredukovanim skupom podataka i algoritamski birao optimalan broj komponenti.

Faktori rizika za samoubistvo su određeni uz pomoć *Random Forest* algoritma. Posmatrana je *Classic* i *Neural* verzija konačnog skupa podataka da bi se dobili najznačajniji faktori rizika na globalnom nivou. Osim ovih faktora na globalnom nivou, posmatrani su i faktori rizika u odnosu na geografsku regiju, religiju, primanja kao i faktori rizika u pojedinim državama. *Random Forest* algoritam je označio, na svetskom nivou, *Fertility rate* kao najuticajniji faktor rizika za samoubistvo u obe verzije skupa podataka. U Istočnoj Aziji i Pacifiku, Južnoj Aziji i Severnoj Americi gustina naseljenosti i rast populacije su predstavljali najznačajnije faktore rizika po algoritmu, dok je za Bliski Istok i Severnu Afriku to bio broj umrlih na 1000 stanovnika. U većini slučajeva pri filtriranju konačnog skupa podataka po primanjima i religiji top deset liste su činile pojedinačne države i regioni, dok su se u slučaju grupisanja po primanjima našle i neke religije. Na kraju su ispitani faktori rizika u pojedinim državama koje su odabrane od strane autora ovog rada. U većini posmatranih država zastupljenost mentalnih poremećaja, kao i zloupotreba alkohola i psihoaktivnih supstanci su predstavljali najbitnije faktore rizika što se slaže sa naučnom literaturom.

2. METODOLOGIJA

Metodologija ovog rada uključuje sledeće oblasti: rad sa skupom podataka, treniranje modela za predikciju i pronalaženje najznačajnijih faktora rizika. U daljem tekstu biće detaljno opisan svaki korak koji je uključen u metodologiju ovog rada po redosledu njihove realizacije počevši od prikupljanja podataka.

2.1. 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. Tabele su prikupljene sa internet stranica institucija kao što su Ujedinjene Nacije zbog kredibiliteta podataka. Ukupno je prikupljeno je 49 tabela koje su analizirane, a izvučena statistika i informacije su pomogle pri daljoj obradi tabela i pronalaženju optimalnog opsega godina.

2.2. Analiza i obrada podataka

Prvo se pristupilo eksplorativnoj analizi prikupljenih skupova podataka. Cilj ovog postupka je bio da se proceni početno stanje tabela, da se iz statističkih podataka svih tabela izračuna optimalni opseg godina za posmatranje koji bi uključivao što je veći opseg godina, a pri tom sadržao što je više moguće podataka. Izabran je opseg od 1990.-2017. koji je sadržao većinu podataka iz odabranih tabela. Problemi koji su se trebali rešiti pre spajanja su popunjavanje nedostajućih podataka, izbacivanje dupliranih podataka i filtriranje tabela na izabran opseg godina. Nedostajući podaci su inicijalno popunjavani sa konstantama (nula, minimumom, maksimumom i srednjom vrednosti iz datog skupa) da bi se tabele mogle iskoristiti za treniranje regresionih algoritama. Odabrani su *Elastic net, XGBoost* i duboke neuronske mreže kao regresioni algoritmi, čija se tačnost nad sve četiri verzije skupova podataka evaluirao. Najbolja verzija iz grupe *Elastic net* i *XGBoost*, kao i *DNN* verzija je iskorišćena za popunjavanje nedostajućih podataka. Originalne tabele sa nedostajućim podacima su propuštane kroz dva regresiona modela i time su nedostajuće vrednosti iz svih skupova podataka bile popunjene. Pre spajanja skupova podataka, sve tabele su filtrirane tako da sadrže samo podatke iz željenog opsega godina.

2.3. Spajanje skupova podataka

Spajanje je vršeno sa *INNER JOIN* metodom, gde su tabele spajane na osnovu zajedničkih atributa - godina i *ISO 3* kôd države. One države koje se nisu našle u svim tabelama su bile izbačene iz skupa podataka. Kreirane su dve verzije konačnog skupa podataka, gde su tabele obrađene sa *Elastic net* ili *XGBoost* algoritmom činile *Classic* verziju, dok su tabele obrađene sa *DNN* činile *Neural* verziju. Razlog za upotrebu ove metode jeste da se osigura kompletnost skupa podataka, tačnije da sve države imaju sve podatke za svaku od godina u odabranom okviru godina, jer treniranje prediktivnih modela zahteva skup podataka bez nedostajućih vrednosti. Time je ukupan broj država spao sa 231 na 142 za *Classic* verziju skupa podataka i 123 za *Neural* verziju skupa podataka. Tabele koje nisu imale nedostajuće vrednosti na početku analize podataka su se spajale i sa tabelama iz *Classic* verzije i sa tabelama iz *Neural* verzije.

2.4. Obeležja spojenog skupa podataka

Nakon kreiranja dve verzije konačnog skupa, broj obeležja u obe verzije je bio 33, a broj država je bio 142 i 123 za *Classic* i *Neural* verziju. Neka od obeležja iz konačnog skupa podataka su: *ISO3* kôd svake od država u skupu, godina, procenat korupcije u državi, ocena poštovanja ljudskih prava, godišnja inflacija, procenat stanovništva koji ima problem sa zloupotrebom alkohola i psihoaktivnih supstanci, dominantna religija, broj dece koji žena rodi tokom svog života i procenat stanovništva koji umre od samopovređivanja. Za predikciju samoubistava bilo je potrebno kreirati modele koji predviđaju poslednje obeležje na osnovu ostalih.

2.5. Enkodovanje string obeležja

Algoritmi za kreiranje prediktivnih modela koji su korišćeni u radu zahtevaju da ulazni skup podataka ima isključivo numeričke vrednosti. U konačnom skupu podataka postoje četiri obeležja koje imaju string vrednost: *ISO 3* kôd, dominantna religija, pripadnost grupi u odnosu na ukupna primanja stanovništva i region. Ova obeležja je bilo neophodno pretvoriti u numeričke vrednosti kroz postupak enkodovanja podataka. Algoritam za enkodovanje podataka koji je korišćen u radu je bio *One-Hot Encoding*. Nakon propuštanja svih verzija skupova podataka kroz algoritam za enkodovanje, dobijeni su skupovi podataka koji su bili spremni za obučavanje predikcionih modela na globalnom nivou. Ovim skupovima se broj kolona povećao na 188 (*Classic* verzija) i 169 (*Neural* verzija) kao posledica enkodovanja string obeležja.

2.6. Redukcija dimenzionalnosti

Usled povećanja broja obeležja, razmatran je uticaj redukcije dimenzionalnosti na tačnost predikcionih modela. Da bi se ispitao ovaj uticaj, bilo je potrebno kreirati skupove podataka sa redukovanim brojem obeležja, a za to su upotrebljeni redukcioni algoritmi *Low Variance Filter* (*LVF*), *Random Forest*, *Principal Component Analysis* (*PCA*), *Factor Analysis* i *Partial Least Squares* (*PLS*). Propuštanjem obe verzije konačnog skupa podataka kroz četiri odabrana redukciona algoritma dobijeno je osam novih skupova podataka. Novodobijeni (redukovani) skupovi podataka zajedno sa dve (neredukovane) verzije konačnog skupa podataka iskorišćeni su za treniranje prediktivnih modela.

2.7. Treniranje prediktivnih modela

Prediktivni algoritmi u radu su korišćeni za dobijanje modela za predikciju samoubistava na globalnom nivou. Ciljno obeležje koje su regresioni modeli trebali da prediktuju jeste stopa samoubistva. Odabrana su tri regresiona algoritma (*Elastic net*, *XGBoost* i *Partial Least Squares*) koja su trenirana nad 10 skupova podataka (dve neredukovane verzije i osam redukovanih). Rezultat treniranja je trideset modela za predikciju stope samoubistava na globalnom nivou, zajedno sa rezultatima validacije ovih modela.

2.8. Optimizacija hiperparametara modela

Optimizacija hiperparametara, tj. parametara modela je rađena sa ciljem dobijanja što boljih rezultata i posvećeno je dosta pažnje empirijskoj optimizaciji istih. Ona je vršena uz pomoć trening skupa i rezultata unakrsne validacije. Parametri algoritama za redukciju dimenzionalnosti optimizovani su empirijski, ali je uzeta u obzir i domenska osnova. Domenska preporuka granične vrednosti kod *Low Variance Filter* algoritma 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 i uzeto je trećina, polovina i dve trećine ukupnog broja obeležja. Parametar *alpha* kod *Elastic net* algoritma je podešen na 0,01, ali isprobane su i vrednosti 0,001, kao i 0,1 i 0,8. Za optimizaciju *XGBoost* algoritma su podešavani *learning_rate* (0,1), *colsample_bytree* (0,3), *max_depth* (5) i *n_estimators* (10) parametri.

2.9. Validacija prediktivnih modela

Za validaciju svih modela su korištene dve metode - srednja kvadratna greška i koeficijent determinacije. Dodatna metoda validacije za *PLS* algoritam je bila unakrsna validacija. Validacija prediktivnih modela je zahtevala da se svi skupovi podataka pre početka treniranja podele 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.

2.10. Pronalaženje najznačajnijih faktora rizika

Cilj pronalaženja najznačajnijih faktora rizika za samoubistvo je bila provera kvaliteta konačnog skupa podataka uz pomoć *Random Forest* algoritma, ali i utvrđivanje da li se kreiran skup podataka i dobijeni modeli za predikciju slažu sa naučno dokazanim činjenicama o faktorima rizika za samoubistvo. Prvo su posmatrani faktori rizika na globalnom nivou tako što su *Random Forest* algoritmu prosleđene *Classic* i *Neural* verzija konačnog skupa podataka. Zatim su posmatrani faktori rizika na nivou geografskih regija, religija, država sa istim prihodima stanovništva i na nivou pojedinih država. Za ove potrebe, obe verzije konačnog skupa podataka su filtrirane na osnovu željenih obeležja – *Region* za geografsku regije, *Dominant religion* za religije, *Income group* za prihode stanovništva i *Country code* za pojedinačne države, a zatim prosleđene *Random Forest* algoritmu. Ovaj algoritam je zatim vraćao listu svih obeležja sortiranu od najuticajnijeg ka najmanje uticajnom obeležju na rezultat predikcije koja je ograničena na deset najznačajnijih obeležja radi lakšeg prikazivanja na grafikonima.

3. REZULTATI

3.1. Rezultati prediktivnih modela

PLS regresioni model treniran sa *PCA* redukovanim *Classic* skupom podataka se pokazao kao najbolji prediktivni model sa koeficijentom determinacije 93,6%. Originalna *Neural* verzija skupa podataka, kao i ona redukovana sa *Factor Analysis* algoritmom je iskorišćena za treniranje *PLS* regresionog modela koji je dao najbolji rezultat od 92,9%. Najlošije rezultate su pokazali *XGBoost* prediktivni modeli. Pokazalo se da su neredukovane verzije u većini slučajeva imale bolje rezultate od svojih redukovanih verzija. Unakrsna validacija je pokazala da je *PLS* algoritam davao najbolje rezultate kada je radio sa neredukovanim skupom podataka i algoritamski birao optimalan broj komponenti (često je taj broj bio ispod 20). Razlike između *Classic* i *Neural* verzije konačnog skupa podataka su vrlo malo uticale na tačnost prediktivnih modela, gde je prosečna razlika bila ispod 4%.

3.2. Najznačajniji faktori rizika

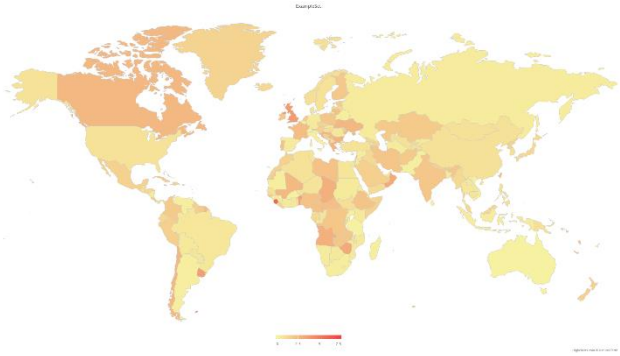
Na globalnom nivou, broj dece koji žena rodi tokom svog života je označen kao najznačajniji faktor rizika. Zdravstveni poremećaji kao što su depresija i bolesti zavisnosti su se našle visoko na globalnoj listi faktora rizika, kao i pojedine države koje imaju visoke stope samoubistava.

Većina faktora rizika koji su se pojavljivali u globalnim listama, nalaze se i u listama za regije. U regijama Istočna Azija i Pacifik, Južna Azija i Severna Amerika gustina naseljenosti i rast populacije su predstavljali najznačajnije faktore rizika.

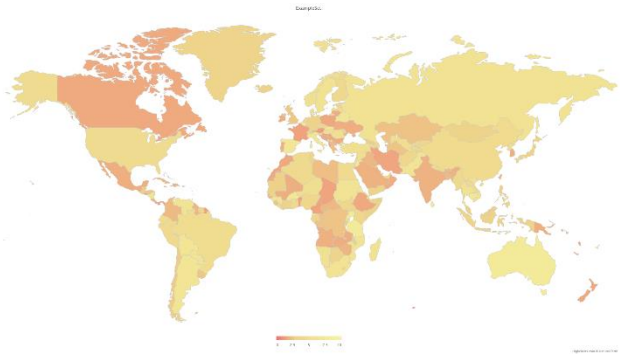
Bliski Istok i Severna Afrika je imala broj umrlih na 1000 stanovnika kao faktor broj jedan za određivanje stope samoubistva. Top deset liste su činile pojedinačne države i regioni kad je vršena filtracija konačnog skupa podataka po primanjima i religiji, a i neke religije su se našle u slučaju grupisanja po primanjima.

Kao na globalnom nivou, većina država su imale zastupljenost mentalnih poremećaja i zloupotrebu alkohola i psihoaktivnih supstanci za najbitnije faktore rizika. Zastupljenost mentalnih poremećaja, depresija, zloupotreba alkohola i psihoaktivnih supstanci su označeni kao najznačajniji faktori rizika u Bugarskoj, Mađarskoj, Litvaniji, Sloveniji, Japanu i Šri Lanki.

Ocena poštovanja ljudskih prava u Hrvatskoj, rast gradskog stanovništva u Crnoj Gori, *Fertility rate* u Severnoj Makedoniji, broj umrlih na 1000 stanovnika u Dominikanskoj Republici, postotak poslodavaca u Indiji i Južnoj Koreji (*Classic* verzije) prepoznati su od strane *Random Forest* algoritma kao najuticajniji na predikciju samoubistva. Korelacija između stope samoubistva (slika 1) i broja dece koji žena rodi tokom svog života (slika 2) može se vizuelno primetiti na mapi sveta.



Slika 1. *Prikaz stope samoubistva na mapi sveta*



Slika 2. *Prikaz obeležja Fertility rate na mapi sveta*

4. ZAKLJUČAK

Korišćenje većeg skupa podataka daje mogućnost dobijanja preciznije predikcije što je možda i ovde bio slučaj. *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* verzijom i *PCA* treniran sa *Classic* verzijom.

Značaj obeležja koji se dobija uz pomoć *Random Forest* algoritma ne pokazuje da li je korelacija pozitivna ili negativna, već samo ukazuje na jačinu iste. Dobijeni najznačajniji faktori rizika samoubistva na svetskom nivou se poklapaju sa faktorima koji su prepoznati u stručnoj literaturi [2]. Države koje danas imaju visoke stope samoubistava, kao što su Južna Koreja, Šri Lanka i Litvanija, nalaze se na top deset listi najznačajnijih faktora rizika [3].

U Indiji i Južnoj Koreji je broj poslodavaca bio bitan faktor što ukazuje na potrebu za finansijskom stabilnosti i nezavisnosti. Osnovna ljudska prava i uređenost države igraju važnu ulogu u prevenciji suicida što ukazuju rezultati Hrvatske, Albanije, Rumunije i Surinama. Prednost ovog rada u odnosu na većinu spomenutih radova jeste korišćenje skupa podataka na globalnom nivou, tako da je uticaj pojedinačnih država mali. Korišćeni skupovi podataka su javno dostupni i publikovani su od strane organizacija kao što su Ujedinjene Nacije.

5. LITERATURA

[1] World Health Organization. Suicide in the world: global health estimates. No. WHO/MSD/MER/19.3. World Health Organization, 2019.

[2] World Health Organization - Suicide: https://www.who.int/news-room/fact-sheets/detail/suicide (pristupljeno u septembru 2021.)

[3] World Health Organization - Global Health Observatory data repository: https://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en (pristupljeno u septembru 2021.)

Kratka biografija:

Boris Bibić rođen je u Subotici 1996. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike - Sistemi za istraživanje i analizu podataka odbranio je 2021. god. kontakt: borisbibic1996@gmail.com