**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 [1] 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. Ovaj rad je pokazao uspešnost *Random Forest* algoritma za predikciju ovakvog tipa problema.

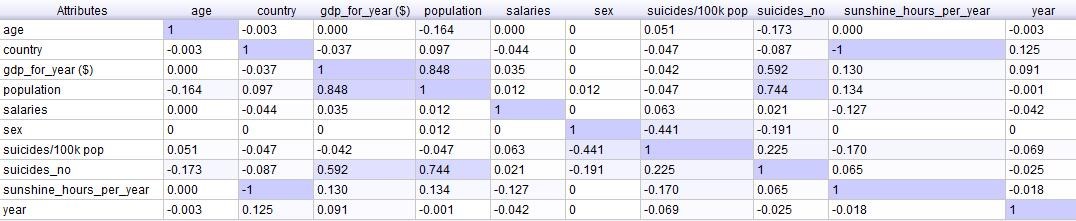
Rad [2] se bavi predikcijom ideja pojedinaca o samoubistvu na osnovu socio-demografskih, fizičkih i psiholoških obeležja. Predikcija se vrši upotrebom *Random Forest* algoritma. Od prvobitnih 47 obeležja, zapaženo je da je model obučen sa 39 karakteristika postigao najveću *Kappa* vrednost. Potom je 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 izbegao *overfitting* i povećala generalizacija modela, korišćena je 10-struka unakrsna validacija. 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 [3] 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. 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 [4] 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.

**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 iz alata Rapid Miner. Rezultat korišćenja ovog operanda je prikazan matrično (slika 1). 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.



Slika 1. Matrica korelacije atributa

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.­­

**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 je 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 primenjen *LabelEncoder* [5]. Korišćena je funkcija *fit\_­­­­­­transform*, 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*.

Sledeća transofrmacija koja je izvršena jeste normalizacija. Normalizacija je primenjena na sve podatke, kako bi se proverilo da li ima uticaj na rezultate predikcije. Normalizacija je izvršena pomoću formule:

gde predstavlja normalizovanu vrednost koja je nalazi u opsegu [0, 1], dok je 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 [5] (eng. *Principal component analysis*) procedura, koja omogućava smanjenje velikog broja atributa. 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*). U tu svrhu je upotrebljena funkcija *cross\_val\_score* iz paketa *sklearn.model\_selection* [5]. Za potrebe unakrsne validacije, skup podataka za trening je podeljen na 5 podskupova koji su dalje korišćeni u procesu validacije. 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 i Lasso regularizaciju. 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.

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.

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).

1. Colin G. Walsh, Jessica D. Ribeiro, Joseph C. Franklin (2017) *Predicting Risk of Suicide Attempts Over Time Through Machine Learning*
2. Seunghyong Ryu, Hyeongrae Lee, Dong-Kyun Lee, and Kyeongwoo Park (2018) *Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population*
3. Sung Man Bae, Seung A Lee, Seung-Hwan Lee (2015) *Prediction by data mining, of suicide attempts in Korean adolescents: a national study*
4. Benjamin Vyssoki, Nestor D. Kapusta, Nicole Praschak-Riede, Georg Dorffner, Matthaeus Willeit (2014) *Direct Effect of Sunshine on Suicide*
5. Scikit-learn: Machine Learning in Python, Pedregosa *et al*., JMLR 12, pp. 2825-2830, 2011.