Predikcija uzroka smrti analizom zdravsvenog kartona korišćenjem metoda mašinskog učenja

Mihajlo Perendija
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Srbija
perendija.e274.2020@uns.ac.rs

Petar Bašić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Srbija
petar.basic@uns.ac.rs

***Apstrakt*—Određivanje rizika od bolesti odnosno uzroka prerane smrti je jedan od ključnih zadataka za prevenciju i borbu protiv mnogih oboljenja. Formiranje sistema za davanje predloga uzroka smrti za neku osobu, na osnovu njene medicinske istorije je upravo tema ovog rada, a njegovo uspešno kreiranje bi pomoglo medicinskom osoblju u ovom vrlo zahtevanom poslu. Medicinska istorija pacijenata je generisana odgovarajućim softverom, zatim detaljno analizirana, obrađena i prebačena u oblik pogodan za modele mašinskog učenja koristeći dva pristupa zasnovana na statističkim metodama i učenju reprezentacije. Korišćeni su modeli mašinskog učenja u režimu višeklasne klasifikacije i u konfiguraciji ansambla modela. Isprobani su modeli veštačkih neuronskih mreža, stabla odlučivanja i model XGBoost. Nakon izvršenih eksperimenata dobijeni su obećavajući rezultati, kojim je dokazan potencijal primenjenih metoda za generisanje reprezentacije heterogenih i istorijskih medicinskih podataka i metoda mašinskog učenja za dalji razvoj i primenu istih u realnom okruženju, nad realnim podacima.**

Ključne reči— uzrok smrti, mašinsko učenje, klasifikacija, elektronski zdravstveni karton, reprezentacija podataka

I. UVOD

Loši uslovi i način vođenja života dovode do preuranjene smrti miliona ljudi svake godine. Poznato je da otkrivanje rizika koji neka osoba ima od oboljevanja od potencijalno smrtonosnog oboljenja može drastično pomoći u prevenciji, ali i budućem lečenju. Zadatak otkrivanja ovih rizika pada prvenstveno na lekare i ostalo medicinsko osoblje. U moderno doba, bilo bi poželjno iskoristiti napredne tehnologije u kombinaciji sa istorijskim podacima i postojećim znanjem, te ponuditi lekarima pomoćni mehanizam za ispunjenje navedenog zadatka. Rizici od oboljenja koji su otkriveni ranije u životu omogućavaju pojedincima da se, uz podršku svog medicinskog tima, skoncentrišu na praćenje, prevenciju i lečenje i tako potencijalno izbegnu preuranjenu smrt. Ovaj rad bavi se predikcijom uzroka smrti nekog pojedinca analizom njegovih osnovnih informacija i informacija o njegovom zdravstvenom stanju tokom života. Predložena metodologija obuhvata detaljnu analizu postojećih informacija o preminulim osobama, određivanje potencijalnih parametara koji su povezani sa uzrokom smrti, zatim preobražavanje odabranih podataka u mašinski obradiv oblik i naposletku kreiranje prediktivnog modela koji bi na osnovu tih parametara nekog živog pojedinca mogao dati najverovatnije uzroke preranog smrtnog ishoda.

Tokom pripreme i rada na ovom projektu postojale su mnogobrojne prepreke, od kojih se izdvaja nemogućnost pribavljanja kompletnih i javno dostupnih podataka. Za to postoje adekvatni razlozi od kojih se izdvaja zaštita podataka o ličnosti, što predstavlja problem i autorima koji se bave rešavanjem sličnih zadataka, te je njihov broj ograničen. Ipak, u narednom poglavlju dat je pregled relevantne literature gde su autori imali pristup realnim podacima, dok je za potrebe ovog rada iskorišćen sintetički generisan skup potrebnih podataka, o čemu će biti reči dalje u tekstu.

Veliki izazov prilikom rada sa podacima iz medicinske istorije ljudi predstavlja njihova heterogenost i međusobna raznolikost, dok značajnu ulogu u otkrivanju nekog oboljenja može imati i vreme nastanka podatka. Dalje u ovom tekstu biće predstavljena dva pristupa za predstavu ovakvih podataka u kombinaciji sa više različitih metoda mašinskog učenja za postizanje krajnjeg cilja.

U narednom poglavlju nalazi se pregled srodnih istraživanja dok su u daljim poglavljima detaljnije izloženi korišćeni podaci, načini njihove pripreme i kreiranja pogodne reprezentacije za primenu metoda mašinskog učenja. U poslednjim poglavljima dat je pregled rezultata primenjenih metoda i izveden zaključak rada.

II. PREGLED RELEVANTNE LITERATURE

Tradicionalno, zadaci predikcije i modelovanja prognoza o budućem zdravstvenom stanju ljudi oslanjali su se na statističke metode sa definisanim algoritmima. Razvoj tehnika mašinskog učenja i mogućnost obrade velikih skupova podataka otvaraju mogućnost za isprobavanje novih i potencijalno revolucionarnih pristupa. Za potrebe ovog rada izdvajaju se tri relevantna istraživanja.

U radu [1] Tong Ruan et al. su predstavili primenu autoenkodera zasnovanog na rekurentnoj neuronskoj mreži za potrebe predstave vremenskih podataka sadržanih u elektronskom medicinskom kartonu. Ovako kreirana predstava podataka pacijenata je korišćena i evaluirana nad više zadataka,

između ostalog, i zadatka predikcije uzroka smrti. Pokazali su da ovakav pristup daje poboljšanje rezultata u odnosu na korišćenje nekih drugih reprezentacija podataka iz elektronskog kartona pacijenta. U ovom radu se koristi ovakav pristup reprezentacije podataka i poredi sa drugim koji će biti opisan u nastavku rada.

U radu [2] Chungsoo Kim et al. akcenat je bio na kreiranju modela mašinskog učenja koji bi na osnovu poslednjeg lekarskog pregleda pacijenta pretpostavio uzrok smrti istog. Dat je predlog modela u obliku ansambla, gde su pojedinačni modeli koji čine taj ansambl trenirani tako da daju pretpostavku da li je uzrok smrti jedan od 8 pojedinačno razmatranih razloga smrti, a te pretpostavke bivaju kombinovane i daju konačnu procenu. Ovaj pristup je evaluiran i u ovom radu ali uz korišćenje drugačijih podataka a njegove performanse su poređene sa jednim modelom koji je u mogućnosti da vrši predikciju nad više uzroka smrti istovremeno.

Stephen F. Weng et al. su u radu [3] prikazali poboljšanje rezultata predikcije preuranjene smrti uz pomoć metoda mašinskog učenja u odnosu na korišćenje standardnih, statističkih i algoritamskih pristupa rešavanju ovog problema predikcije. Korišćen je drugačiji pristup reprezentacije vremenskih podataka u odnosu na onaj koji je korišćen u ovom radu ali su prikazane performanse korišćenja modela neuronskih mreža i stabala odlučivanja što je autore ovog rada navelo ka izboru tih modela.

III. Opis skupa podataka

Za ispunjenje ciljeva ovog projekta potrebno je posedovati podatke o zdravstvenim kartonima preminulih osoba. Za pristup ovim podacima, a usled njihove visoke osetljivosti, potrebno je posedovanje posebnih dozvola. U nedostatku potrebnih dozvola i nepostojanju adekvatnih javno dostupnih podataka, iskorišćen je softverski alat za generisanje sintetičkih podataka *Synthea* [4], čijom upotrebom su dobijeni podaci za 23000 preminulih osoba. Iako se podaci generisani na ovaj način ne odnose na realne osobe, njihova validnost u smislu medicinske tačnosti i sličnosti sa pravim podacima prikazana je u radovima [5] i [6].

Podaci generisani upotrebom *Synthea* softvera predstavljaju zdravstvene kartone pacijenata koji sadrže: statičke osnovne informacije o pacijentu (ime, godina rođenja i smrti, pol, bračni status, rasa, socio-ekonomski podaci, prebivalište), informacije o pregledima (datum održavanja, tip pregleda (razlog obavljanja), dijagnoza, razlog smrti – ukoliko je to razlog obavljanja pregleda), zapažanja i merenja medicinskog osoblja na pregledima (visina, težina, pritisak, puls, izveštaji sa laboratorijskog nalaza itd) i informacije o stanju pacijenta (podaci o bolestima i to identifikator bolesti i vreme trajanja oboljenja). Zabeležene su i informacije o propisanim terapijama i lekovima. Pored ovih, pokazaće se i najbitnijih, postoje i podaci o alergijama, imunizaciji i institucijama u kojima je pacijent lečen. Kako je cilj projekta određivanje verovatnoće smrtnog ishoda usled nekog oboljenja, glavno obeležje koje služi u svrhe predikcije predstavlja uzrok, odnosno glavno oboljenje koje je dovelo do smrti pacijenta. Iako generisani, ovi podaci predstavljaju realnu sliku odnosa i razlika u broju osoba preminulih od različitih oboljenja, što se manifestuje nebalansiranim skupom podataka.

Inicijalno generisanje sintetičkih podataka uključivalo je i grupe podataka koje su, nakon kratkog vizuelnog pregleda, unapred zanemarene. Ove grupe obuhvataju podatke o zdravstvenom osiguranju pacijenata, osiguravajućim kućama, bolnicama, novčanim transakcijama koje su izvršene, potrošenim zalihama po bolnici, alergijama, imunizaciji i podatke o izvršenim radiološkim pregledima. Uzimajući u obzir rezultate rada [3] zaključeno je da je značaj navedenih podataka za konkretan zadatak zanemarljiv, dok bi njihovo uključivanje doprinelo kompleksnosti i usporilo eksperimente.

Početnom analizom zaključeno je da postoje podaci o osobama čija smrt nije uzrokovana u potpunosti prirodnim putem. Nasilnu, smrt nastalu nakon nagle povrede i smrt koja je rezultat korišćenja opojnih supstanci nije moguće zdravorazumski dovesti u vezu sa medicinskom istorijom pacijenta, stoga podaci o pacijentima čiji uzrok smrti spada u ove grupe nisu adekvatni za zadatak ovog rada. Većina ovakvih podataka nije uzeta u obzir, dok su podaci koji se vezuju za nekoliko naglih fizičkih povreda uključeni u dalju obradu radi dokazivanja ove pretpostavke. Takođe, izbačeni su i podaci čiji uzrok smrti nije bio poznat (nedostajuća vrednost), čime se došlo do ~21000 podataka.

Oslanjajući se na rezultate rada [3] izvršena je eksplorativna analiza podataka u cilju pronalaska obrazaca i obeležja koja imaju uticaja na ciljno obeležje tj. uzrok smrti. Utvrđena je izražena zastupljenost različitih oboljenja, terapija, lekova i vrednosti izvršenih merenja tokom života koji se vezuju za pojedinačni uzrok smrti.

Iako je u radu [3] naglašen uticaj socioekonomskih faktora na krajnji uzrok prerane smrti, u podacima korišćenim u ovom istraživanju nije pronađena značajna korelacija, što se može pripisati težnjom kreatora *Synthea* softvera za tačnim modelovanjem medicinske istorije u opštem slučaju. Dodatno, programski generisani podaci ne sadrže podatke o porodičnoj istoriji oboljenja niti potpune podatke o životnim navikama, ali će se pokazati da i redukovani skup podataka ima potencijal za dostizanje dobrih rezultata.

Ukoliko se osvrnemo na starost populacije u trenutku smrti, može se primetiti visoka zastupljenost osoba ispod 12 godina (~18 %), kao i zastupljenost jednog oboljenja (iznenadni zastoj srca -*eng. sudden cardiac arrest*) među ovom populacijom. Iz ovih razloga rezultate koji se tiču ove starosne grupe treba uzeti sa rezervom.

Detaljnijim pregledom podataka primećena je, očekivano, drastično veća povezanost zapisa iz zdravstvenog kartona nastalih neposredno pred smrt sa uzrokom smrti. Radi uvođenja realnosti u istraživanje, u dalje razmatranje su uzeti isključivo podaci nastali do ~4 meseca pred smrt. Na ovaj način izbegnuto je donošenje očiglednih zaključaka (kao na primer, uzrok poslednjih nekoliko pregleda pacijenta su prijem u bolnicu zbog šloga, što je i uzrok smrti).

Nakon navedenih analiza utvrđeno je da postoje ljudski uočljivi i objašnjivi obrasci koji povezuju istorijske podatke iz zdravstvenih kartona pacijenata sa njihovim uzrokom smrti.

IV. METODOLOGIJA

Nakon generisanja, analize i obrade podataka potrebno ih je prevesti u oblik pogodan za mašinsku obradu, za šta su isprobana dva nezavisna pristupa odnosno metode vektorizacije. Naposletku, primenjeno je više poznatijih metoda mašinskog učenja, koristeći prethodno vektorizovane podatke, a za zadatak klasifikacije medicinskih podataka pacijenta u zavisnosti od uzroka smrti.

A. Metode vektorizacije

Prvi pristup obuhvata obučavanje autoenkoder [7] modela nad pripremljenim podacima, zatim izdvajanje enkoder dela već istreniranog modela i njegovo korišćenje za kodiranje odnosno kreiranje vektorske reprezentacije podataka.

Zdravstveni karton pacijenta sadrži sekvence podataka zabeleženih u različitim vremenskim trenucima. Cilj je pretvoriti ovu sekvencu događaja u jednu vektorsku reprezentaciju pacijenta, uz zadržavanje vremenske dimenzije. U radu [1] se upravo predlaže slična metoda učenja reprezentacije pacijentovih podataka koja koristi autoenkoder zasnovan na rekurentnim neuronskim mrežama (RNN) za enkodiranje vremenskih serija.

Za rešavanje problema predikcije vremenskih serija se često koriste RNN zbog njihove sposobnosti pamćenja istorijskih informacija, a usled postojanja problema eksplozije i nestajućeg gradijenta razvijen je LSTM model (*long short-term memory*), odnosno njegova pojednostavljena verzija GRU (*gated recurrent unit*), koji se u ovom radu koristi tokom istraživanja.

Priprema podataka za enkodiranje GRU enkoder modelom, pored inicijalne analize i obrade, obuhvata:

- Konverziju pojedinačnih događaja iz zdravstvenog kartona u *multi-hot* vektore, što se može objasniti na primeru. Ukoliko posmatramo grupu podataka o prepisanim lekovima, jedan zapis je dimenzije 4 i sadrži identifikator pacijenta, identifikator pregleda na kom je lek prepisan, opis i kod leka i vreme prepisivanja leka. Ukoliko je na jednom pregledu prepisano više od jednog leka, postojaće i više zapisa o istom događaju u ovoj grupi. Preoblikovanjem se dobija *multi-hot* vektor dimenzije 217, odnosno sadrži identifikator pacijenta i konkretnog događaja (pregleda) i indikatore o prepisanim lekovima na odgovarajućim mestima. Dodatne dimenzije čine svi mogući prepisani lekovi koji su sadržani u podacima. Konkretan, pojednostavljeni, primer prikazan je na slici 1. Izuzetak predstavljaju podaci o zapažanjima i merenjima gde je umesto indikatora postojanja postavljena stvarna vrednost merenja.

Patient ID	Encounter ID	Medication
0dddc561	38aafe98	piperacillin
0dddc561	38aafe98	vancomycin
0dddc561	38aafe98	Nitrofurantoin

Patient ID	Encounter ID	piperacillin	Naproxen	…	vancomycin	Nitrofurantoin
0dddc561	38aafe98	1	0	…	1	1

Slika 1 Konverzija podataka u multi hot vektore

- Udruživanje pojedinačnih grupa *multi-hot* vektora u jedan krajnji *multi-hot* vektor za konkretan događaj nekog pacijenta. Slika 2 predstavlja jedan od koraka prilikom udruživanja grupa podataka.

Observations

Patient ID	Encounter ID	Diastolic Blood Pressure	Cholesterol	…
0dddc561	d2cfbdae	85.0	165.3	…
0dddc561	49b59550	90.0	169.2	…

+

Medications

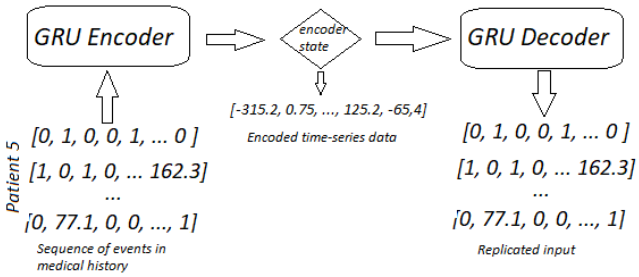
Patient ID	Encounter ID	piperacillin	…	Estrostep
0dddc561	d2cfbdae	0	…	1
0dddc561	49b59550	1	…	0

Patient ID	Encounter ID	piperacillin	…	Estrostep	Cholesterol	…	Dias. Blood Preasure
0d..	d2..	0	…	1	165.3	…	85.0
0d..	49..	1	…	0	169.2	…	90.0

Slika 2 Proces udruživanja vektora iz različitih grupa u jedan krajnji vektor

Podaci jednog pacijenta sada su predstavljeni kao sekvenca konvertovanih i udruženih vektora.

Sledeći korak je obučavanje autoenkoder modela, čiji su ulazi, sekvence vektora pacijenata, replikovani na izlaz. Iz razloga što su sekvence varijabilne dužine, izvršeno je popunjavanje, odnosno njihovo skraćivanje (*eng. padding*). Na slici 3 prikazana je skica ovog metoda nakon izvršenog obučavanja autoenkoder modela. Izlazi iz enkoder dela modela, na slici prikazani kao *encoder state* (stanje enkodera u izlaznom sloju), zapravo prestavljaju rezultat ovog metoda – kodirani zdravstveni karton pacijenta.



Slika 3 Skica rada obučenog *autoenkoder* modela

Drugi pristup vektorizaciji nastao je tokom rada na prvom pristupu. Usled nedostatka kompjuterskih resursa bilo je potrebno znatno smanjiti dimenzionalnost podataka i to izbacivanjem grupe podataka o zapažanjima i merenjima i ograničavanjem broja vektora u sekvenci na 30. Postojala je pretpostavka da se ovakvim skraćivanjem podataka gube značajne informacije, te se paralelno pribeglo razvijanju jednostavnijeg metoda vektorizacije.

Jednostavnost drugog pristupa ogleda se u odsustvu bilo kakvog modela mašinskog učenja u procesu vektorizacije. Vektorizacija je izvršena isključivo preoblikovanjem, agregacijom i ostalim vrstama obrade podataka.

Proces vektorizacije ovim metodom je u osnovi identičan prethodnom pristupu, do trenutka dobijanja sekvenci vektora za svakog pacijenta. Razlike su vidljive u:

- procesu konverzije grupe podataka o zapažanjima i merenjima. Ako podatke posmatramo u tabelarnom obliku, ovaj pristup bi za svaki jedinstveni tip merenja koji se javlja kao vrednost u početnoj tabeli, kreirao K kolona u rezultujućoj tabeli, gde je K parametar određen eksperimentalnim putem. Parametar K predstavlja broj podela početnog merenja na podgrupe, odnosno raspone vrednosti tog merenja. Zgodan primer prikazan je na slici 4 gde se može videti konverzija merenja telesne mase pacijenta u 5 raspona težina. Pacijentova rezultujuća tabela za podatke o merenjima sada sadrži *multi-hot* vektore.

  Konkretni rasponi, odnosno granice grupa u koje se neko merenje deli su unapred sračunati. U obzir su uzete sortirane vrednosti za konkretno merenje za celu populaciju. Granice su onda formirane tako da svaka grupa, na celokupnoj populaciji, ima isti broj merenja.

Patient ID	Encounter ID	Observation	Value
6772742f	58eac888	Body Weight	64.0

⇩

Patient ID	Encounter ID	Weight <45	Weight >45<74	Weight >74<91	Weight >91<123	Weight >123
6772..	58eac888	0	1	0	0	0

Slika 4 Primer konverzije podataka o merenjima.

- u načinu unošenja vremenske dimenzije u podatke. Prethodni metod se oslanja na osobine RNN za pamćenje istorijskih informacija. U ovom metodu se predlaže pojednostavljenje ovog zadatka oslanjanjem na pretpostavku o manjoj važnosti najstarijih podataka. Na ovaj način se vrednosti podataka koji su nastali najranije u životu pacijenta „kažnjavaju“ najviše, tj. najviše im se umanjuje značaj. Na početku procesa se za svaki podatak određuje udaljenost njegovog nastanka od trenutka smrti. U zavisnosti od ove vrednosti se prilikom kreiranja multi hot vektora, umesto jednostavnog indikatora, postavlja težinski faktor. Na slici 5 prikazan je pojednostavljeni postupak kreiranja *multi-hot* vektora, sa umanjivanjem značaja starim podacima na primeru telesne mase. Vrednosti težinskog faktora dobijene su nakon nekoliko izvršenih eksperimenata, kao one sa kojima su dobijeni najbolji krajnji rezultati. Podacima starijim od 30 i 20 godina značaj je umanjen za 90 % odnosno 60 % respektivno, između 5 i 20 godina 40 %, između 1 i 5 godina 10 %, dok najskorijim podacima značaj nije umanjivan.

Patient ID	Encounter ID	Date	Distance from death date [yrs]	Observation	Value [kg]
6772..	58eac888	1990-03-19	4.6 {0.9}	weight	85
6772..	14jag887	1985-03-13	9.6 {0.6}	weight	115
6772..	74ea12g6	1964-05-12	20.4 {0.4}	weight	40
6772..	Zp96c452	1962-07-15	22.2 {0.4}	weight	35

⇩

Patient ID	Weight <45	Weight >45<74	Weight >74<91	Weight >91<123	Weight >123	…
6772..	0.4 + 0.4	0	0.9	0.6	0	…

Slika 5 Kreiranje *multi-hot* vektora sa agregacijom i umanjivanjem značaja starim podacima.

Krajnji vektor kojim se predstavlja jedan pacijent dobijen je agregacijom sekvenci podataka pacijenta za svaki obavljeni pregled. Jedna vrednost iz vektora predstavlja normalizovanu sumu penalizovanih vrednosti za odgovarajući podatak iz zdravstvenog kartona. Na primeru sa slike 5 može se videti isečak krajnjeg vektora pacijenta *6772..* kome su dva puta zabeleženi podaci o telesnoj masi ispod 45kg stariji od 20 godina (težinski faktor 0.4), te krajnji rezultat za taj parametar iznosi 0.8, pre normalizacije.

B. Metode klalsifikacije

Problem koji se rešava u ovom radu se svodi na problem klasifikacije. Na osnovu kodiranih podataka pacijenta, dobijenih prethodno opisanim metodama vektorizacije, potrebno je doneti pretpostavku o mogućim uzrocima smrti.

U slučaju kada imamo više klasa, odnosno uzroka smrti, postoje dva arhitekturalna pristupa pri kreiranju klasifikacionog modela. Prvi pristup jeste kreiranje *multiclass* modela, na čijem izlazu se nalazi jedna izabrana klasa od N mogućih (odnosno verovatnoće za svaku klasu pojedinačno), a drugi pristup je kreiranje ansambla od N modela gde se za svaku klasu obučava zaseban model koji nam govori da li taj podatak pripada toj klasi ili ne.

Postoji mnoštvo modela mašinskog učenja koji se mogu iskoristiti za potrebe klasifikacije. Na izbor je uticala brzina kreiranja, performanse, lakoća optimizacije i preporuke relevantne literature. U ovom radu su korišćeni sledeći modeli: Potpuno povezana neuronska mreža (*eng. fully connected neural network*), stablo odlučivanja (*eng. decision tree*) i XGBoost (*gradient boosted trees*). Svaki od ovih modela je korišćen u obe arhitekture klasifikacionih modela a davali su uporedive rezultate. Određeni modeli su bolje prepoznavali neke uzroke smrti u odnosu na druge modele, i obrnuto. Ipak, veća pažnja je data neuronskim mrežama i stablima odlučivanja iz razloga što je njihovo obučavanje bilo brže te se moglo pristupiti optimizaciji rezultata.

V. REZULTATI I DISKUSIJA

Koristeći prethodno opisane prečišćene podatke i primenjujući navedene metode vektorizacije i klasifikacije izvršeni su brojni eksperimenti u potrazi za najboljim rešenjem zadatka ovog rada. Izvršeni eksperimenti mogu se generalno podeliti u dve grupe. Ono što izdvaja prvu grupu od druge jeste primenjena metoda vektorizacije podataka.

Prilikom klasifikacije podaci su u svakom od eksperimenata podeljeni na trening (64%), validacioni (16%) i test skup (20%). Do navedenih odnosa medju podeljenim podacima se došlo empirijski jer se utvrdilo da daju dobar balans potreban za treniranje, validaciju i evaluaciju. Za evaluaciju rezultata nad pojedinačnim klasama korišćene mere preciznosti, odziva i F1 mera. Na nivou svih klasa zajedno korišćene su iste mere sračunate *micro* i *macro average* metodama kojima se može zaključiti odnos rezultata modela za najzastupljenije i najmanje zastupljene klase.

A. Eksperimenti sa vektorizacijom podataka korišćenjem autoenkoder modela

Generalna procedura izvršena u eksperimentima iz ove grupe obuhvata: pripremu podataka izbacivanjem neželjenih i filtriranjem na osnovu vremena nastanka u odnosu na trenutak smrti, kreiranje sekvenci *multi-hot* vektora, obučavanje i optimizacija autoenkoder modela, korišćenje enkoder dela obučenog autoenkodera za kodiranje podataka i obučavanje višeslojne neuronske mreže za zadatak višeklasne klasifikacije. Glavne razlike između eksperimenata iz ove grupe su u količini i vremenskom nastanku korišćenih podataka.

Važno je napomenuti da je primena konkretne metode vektorizacije vrlo zahtevna u pogledu računarskih resursa. Iz ovog razloga je u ovoj grupaciji eksperimenata korišćen redukovan skup podataka, dobijen odbacivanjem grupe podataka o zapažanjima i merenjima, koja je ujedno i najveća, dok su u obzir uzeti podaci od ~18000 osoba (što je smanjenje od ~3000 u odnosu na originalne podatke).

Za popunjavanje sekvenci podataka (*eng. padding*) koje je dalo najbolje rezultate korišćena je veličina sekvence 30, dok su, ukoliko ih je bilo više od 30, odsečene sekvence sa najstarijim podacima. Svaki vektor u sekvenci sačinjen je od 789 vrednosti.

Arhitektura izabranog autoenkoder modela je jednostavna, sa jednim ulaznim potpuno povezanim i jednim GRU slojem u enkoder delu i jednim GRU i jednim izlaznim potpuno povezanim slojem u dekoder delu. Usložnjavanje modela nije dalo drastično bolje rezultate dok je značajno uticalo na korišćenje resursa i vreme obučavanja. Na osnovu preporuka iz rada [1] i u skladu sa resursnim mogućnostima, određeno je da izlazni sloj enkoder dela modela bude dužine 256, što praktično znači da su podaci jednog pacijenta dužine 789 i širine 30 kodirani u jedan vektor dužine 256.

Odnos performansi i tačnosti klasifikacije bio je najbolji za neuronsku mrežu sa: tri skrivena sloja (redom veličine 1024, 128 i 64 čvorova), *batch* normalizacijom i *dropout*-om (20%) između slojeva, „selu“ aktivacionom funkcijom na skrivenim slojevima, dok je na izlaznom sloju korišćena „softmax“ aktivacija. Upotrebljena je „Binary Cross Entropy“ *loss* funkcija.

Izvršeni su eksperimenti sa podacima nastalim do ~1.7 i do ~0.5 godina od trenutka smrti. Dodatno, radi potpunog dokazivanja upotrebljivosti primenjene metode vektorizacije, izvršen je i eksperiment samo nad podacima nastalim neposredno pred smrt tj. u poslednjih 1.7 godina. Rezultati su prikazani u tabeli 1. Primećuju se izrazito dobri rezultati eksperimenta u kom su korišćeni najskoriji podaci iz medicinske istorije pacijenta. Iako replikovanje ovog eksperimenta na stvarnom primeru nema realnu upotrebnu vrednost (osoba je blizu kraja života), dokazano je da kodiranje vremenskih podataka na ovaj način ima smisla. Realniji rezultati dobijeni su korišćenjem starijih podataka, dok je primetan trend rasta vrednosti svih mera sa povećanjem količine podataka i unošenjem podataka nastalih vremenski bliže smrti.

Korišćeni podaci	Mera	Micro avg			Macro avg		
		Precission	Recall	F1	Precission	Recall	F1
Do ~1.7 god od smrti		0.60	0.33	0.43	0.27	0.17	0.18
Do ~0.5 god od smrti		0.58	0.56	0.57	0.32	0.30	0.30
Od ~1.7 do smrti		0.97	0.96	0.97	0.82	0.70	0.74

Tabela 1 Rezultati eksperimenata korišćenjem autoenkoder modela za vektorizaciju podataka i neuronske mreže za klasifikaciju.

Poređenjem vrednosti za *micro* i *macro average* zaključuje se da primenjena metodologija daje bolje rezultate za više zastupljene klase, odnosno da se povećanjem količine podataka za obučavanje poboljšavaju i rezultati.

Usled ograničene količine dostupnih resursa i nepostojanja realnih podataka, rezultati se samo delimično slažu sa zaključcima iz rada [1], u kom je sličan metod vektorizacije dao najbolje rezultate.

B. Eksperimenti sa statističkom reprezentacijom podataka

Procedura eksperimenata iz ove grupe razlikuje se od prethodne u postupku vektorizacije podataka, što je ujedno pojednostavilo postupak, omogućilo izučavanje više različitih klasifikacionih modela i dozvolilo korišćenje neredukovanog skupa podataka (uključujući zapažanja i merenja).

Pristup vektorizaciji podataka u ovim eksperimentima direktnije predstavlja pojedinačne pacijente, u smislu da ako pacijentu nikada nije dijagnostifikovana određena bolest, na mestu u vektoru koji predstavlja tu dijagnozu će se nalaziti broj 0. U slučaju da jeste, na tom mestu će se nalaziti neki broj koji je između 0 i 1 i koji je dobijen ranije predstavljenim pristupom. S tim u vezi, pretpostavka je da bi modeli kao što su stabla odlučivanja i XGBoost mogli iz takvih slučajeva doneti bolje zaključke.

Prilikom vršenja eksperimenata došlo se do optimalnih parametara svih navedenih modela, negde programskim putem a negde ručnim ponavljanjem eskperimenata i u nastavku su opisani ti procesi za svaki model.

Eksperimenti sa neuronskom mrežom su vršeni ručno i započeti su sa jednostavnom mrežom koja je davala zadovoljavajuće rezultate. Ubrzo je uviđeno da povećavanje broja skrivenih slojeva kao i broja čvorova u prvim slojevima drastično povećava rezultate. Pored toga, uvođenje batch normalizacije je veoma pozitivno uticalo na odziv modela.

Na ovaj način se došlo do optimalne arhitekture neuronske mreže koja se sastoji od:

- 3 skrivena sloja u obliku levka (2046, 516, 128 čvorova po slojevima)
- *Batch* normalizacije između svakog sloja

- "Selu" aktivacionom funkcijom na skrivenim slojevima
- *Dropout* od 20% između svakog sloja
- „Categorical Crossentropy“ *loss*
- „Softmax“ aktivacionom funkcijom na izlaznom sloju
- „Adam“ optimizator

Ova arhitektura je korišćena u oba pristupa i u ansamblu i u *multiclass* pristupu jer su prethodno navedena zapažanja imala sličan uticaj u oba slučaja.

Stabla odlučivanja, korišćena u arhitekturi ansambla, su se pokazala veoma brzim za obučavanje i davala su uporedive rezultate sa neuronskom mrežom te se pribeglo programskoj (automatskoj) optimizaciji parametara. Na taj način se došlo do sledećih parametara:

- Kriterijum mere kvaliteta podele: "entropy"
- Maksimalna dubina stabla: 3
- Minimalni broj primera potrebnih za podelu čvora: 2
- Minimalni broj primera potrebnih da čvor bude list: 1

Pretpostavka o tome da bi ovaj model mogao dati dobre rezultate s obzirom na podatke sa kojima radi se ispostavila tačnom. Naime, za određene uzroke smrti je davao dosta bolje pretpostavke, pogotovu gledajući odziv modela, što se može videti na primerima iz tabele 2.

	F1 NN	Recall NN	F1 Decision tree	Recall Decision tree
maligni tumor debelog creva	0.77	0.78	0.94	1.0
infarkt miokarda	0.67	0.62	0.80	0.87

Tabela 2 Poređenje rezultata klasifikacije sa neuronskom mrežom i stablom odlučivanja primenjenim u arhitekturi ansambla

Ipak, poredeći performanse za sve razloge smrti objedinjeno, rezultati su uporedivi sa neuronskom mrežom.

Na osnovu rezultata modela stabla odlučivanja dalo se za pretpostaviti da će model XGBoost dati još bolje rezultate. Ovom modelu je trebalo najviše vremena za obučavanje te je ručno isproban mali broj kombinacija parametara, dok je korišćena *multiclass* arhitektura. Kao najbolja kombinacija istakla se ona koja ima sve podrazumevane parametre osim: 'objective':'multi:softmax' i 'max_depth':10.

U tabeli 3 se može videti prikaz i poređenje najboljih rezultata modela i to posmatrajući sve uzroke smrti.

	macro f1	macro recall
nn multiclass	0.55	0.54
nn ansambl	0.58	0.58
decision tree ansambl	0.56	0.56
XGBoost multiclass	0.67	0.67

Tabela 3 Poređenje najboljih rezultata eksperimenata sa statističkim pristupom vektorizaciji podataka

Težnja da se pronađe model koji bi radio dobro za sve uzroke smrti navodi nas na evaluaciju ovih modela macro merom. U prilog tome ide i činjenica da nebalansiranost podataka nije znatno uticala na rezultate kao što je to slučaj kod eksperimenata iz prve grupe. Zapaženo je da kada se posmatra svaka bolest pojedinačno, performanse ovih modela variraju te bi kombinacija ovih modela i kreiranje hibridnog klasifikacionog modela bila najbolja opcija.

Za potrebe poređenja sa prethodnim eksperimentima, koji koriste vektorizaciju uz pomoć autoenkodera, izvršen je eksperiment nad redukovanim podacima. Iz podataka su izbačeni podaci iz grupe zapažanja i merenja i uzeti su obzir samo oni nastali do 0.5 godina od smrti osobe. Primenjena je statistička vektorizacija, a za klasifikaciju je korišćena prethodno opisana neuronska mreža.

DISKUSIJA

Vršenjem opisanih eksperimenata došlo se do zaključka da ne postoji jedna kombinacija vektorizacije i klasifikacije koja daje bolje rezultate od svih ostalih i to za sve uzroke smrti. Ipak, izdvaja se pristup gde se koristi reprezentacija podataka dobijena statisičkom vektorizacijom i korišćenjem *multiclass* XGBoost modela.

Measures / Reason of death	support	precision	recall	f1-score
Alzheimer's disease (disorder)	6	0.5	0.5	0.5
Burn injury(morphologic abnormality)	33	0.0	0.0	0.0
Cardiac Arrest	86	0.38	0.06	0.1
Chronic congestive heart failure (disorder)	175	0.94	1.00	0.97
Chronic obstructive bronchitis (disorder)	56	0.88	0.8	0.84
Concussion injury of brain	95	0.31	0.05	0.09
Death due to acute respiratory failure	6	1.00	0.83	0.91
Familial Alzheimer's disease of early onset (disorder)	9	0.56	0.56	0.56
Fracture of the vertebral column with spinal cord injury	43	0.0	0.0	0.0
Malignant neoplasm of breast (disorder)	38	0.97	0.76	0.85
Myocardial Infarction	304	0.77	0.87	0.81
Neoplasm of prostate	13	0.75	0.69	0.72
Non-small cell lung cancer (disorder)	40	0.95	0.97	0.96
Overlapping malignant neoplasm of colon	14	0.93	1.00	0.97
Pneumonia	16	0.4	0.5	0.44
Pulmonary emphysema (disorder)	50	0.67	0.92	0.77
Small cell carcinoma of lung (disorder)	6	1.0	1.0	1.0
Stroke	94	0.0	0.0	0.0

Tabela 4 Isečak rezultata najboljeg rešenja (vektorizacija statističkim metodatama i XGBoost modelom multiclass klasifikacije)

Iz tabele 4 se može primetiti da klasifikacioni model daje izuzetno dobre rezultate čak i za one klase kod kojih postoji mali broj primera, kao na primer "Small cell carcinoma of lung (disorder)" i "Death due to acute respiratory failure". To se može objasniti time da se u medicinskom kartonu nalaze stavke koje su blisko povezane sa određenim postupcima koji se obavljaju u slučaju dijagnoze ovih bolesti, npr. hemoterapija.

Sa druge strane, postoje i razlozi smrti koje model nije uopšte uspeo da predvidi ili je davao loše rezultate. Primeri toga su "Fracture of the vertebral column with spinal cord injury" i "Stroke". U oba slučaja, uzroci smrti nisu posledica nečega što se moglo pronaći u istoriji pacijenta već su posledica nečega što se dogodilo neposredno pred smrt pacijenta, a naročito smrt nastala usled iznenadne fizičke povrede. Ovakvo ponašanje je očekivano i ide u prilog činjenici da se za prava oboljenja mogu naći obrasci u podacima koji usmeravaju modele ka tačnim rezultatima.

Analizom rezultata pojedinačnih modela stabala odlučivanja uočeno je da pri donošenju odluke najviše utiču upravo podaci o zapažanjima i merenjima. U pristupu sa

vektorizacijom uz pomoć autoenkodera, zbog nedostatka resursa, nisu korišćeni ovi podaci te rezultati ta dva pristupa nisu direktno uporedivi, ali se poređenje može vršiti na eksperimentima izvršenim nad redukovanim skupom podataka. Iz tabele 5 može se videti vrlo mala razlika između rezultata dva pristupa, što se može pripisati razlici u broju korišćenih podataka. Zaključuje se da su pristupi uporedivi i da se način vektorizacije uz pomoć autoenkoder modela pokazao podjednako povoljan, a pretpostavlja se i tendencija poboljšanja rezultata u slučaju neredukovanja skupa podataka, a uzimajući u obzir eksperimente iz drugog pristupa. Ipak, njegovim korišćenjem se gubi mogućnost interpretabilnosti u odnosu na statističke metode.

Metodologija	Mera	Micro avg			Macro avg		
		Precission	Recall	F1	Precission	Recall	F1
~18000 osoba & do ~0.5 god od smrti & vektorizacija sa autoenkoder modelom & NN		0.58	0.56	0.57	0.32	0.30	0.30
~21000 osoba & do ~0.5 god od smrti & vektorizacija statističkom metodom & NN		0.68	0.60	0.64	0.52	0.41	0.45

Tabela 5 Poređenje rezultata eksperimenata sa korišćenjem statistički zasnovane vektorizacije i vektorizacije sa autoenkoder modelom

Upravo interpretabilnost modela stabla odlučivanja a donekle i XGBoost modela je njhova najveća prednost s obzirom da se radi o problemu gde je objašnjenje rezultata klasifikacije (pretpostavke) možda i bitnije od samih rezultata.

VI. ZAKLJUČAK

U ovom radu prikazana su rešenja zadatka otkrivanja uzroka prevremene smrti čoveka na osnovu podataka iz zdravstvenog kartona, a uzimajući u obzir vreme njihovog nastanka. Podaci su generisani softverskim alatom *Synthea* a zatim analizirani, isfiltrirani i obrađeni. Priprema podataka za obučavanje modela mašinskog učenja tekla je u dva pravca i to generisanjem vektorske reprezentacije uz pomoć autoenkoder modela tj. učenjem reprezentacije podataka i metodom vektorizacije zasnovanom na statističkoj obradi. Generisane reprezentacije podataka korišćene su za obučavanje više različitih modela mašinskog učenja za zadatak klasifikacije pacijenata na osnovu uzroka smrti u konfiguraciji binarne klasifikacije za svako oboljenja pojedinačno i u *multiclass* režimu.

Sistem koji na osnovu medicinske istorije pacijenta može da da predlog uzroka njegove smrti bio bi značajno pomoćno sredstvo medicinskom osoblju prilikom dijagnostifikovanja, otkrivanja i procene rizika od oboljenja, što može poboljšati prevenciju i olakšati oporavak.

Rezultati metodologije kojom se došlo do najboljih rezultata, ali i ostalih, dokazuju upotrebnu vrednost primenjenih metoda vektorizacije i modela mašinskog učenja za konkretan zadatak imajući u vidu kompleksnost i heterogenost podataka, koji se na prikazane načine mogu efektivno predstaviti u obliku pogodnom za mašinsku obradu. S tim u vezi uočena je i važnost prethodne analize, obrade i pripreme sirovih podataka. Koliko je poznato autorima, u trenutku pisanja rada nisu postojala istraživanja koja su u obzir uzimala tako široku paletu različitih vrsta podataka u kombinaciji sa konkretnim metodama vektorizacije i klasifikacije.

Dobijena rešenja, pak, nisu u stanju da otkrivaju uzroke smrti od svakog oboljenja podjednako. Primetni su i lošiji rezultati ukoliko za pojedinog pacijenta ne postoji kompletna medicinska istorija ili se radi o mlađoj osobi, koja nema veliku količinu podataka. Na osnovu ovoga se može zaključiti važnost obavljanja redovnih medicinskih analiza i pregleda.

Praktična primena prikazanih rešenja, zbog brojnih ograničenja, bila bi svedena na korišćenje istraživanja kao podloge za dalji razvoj i kao izvor pokazatelja uspešnosti prikazanih metodologija.

Prvi korak u budućem istraživanju bio bi prikupljanje stvarnih podataka i to u mnogo većim količinama kako bi se iskazala verodostojnost rezultata. U slučaju realnih podataka, predlaže se detaljnija analiza i obrada podataka radi uočavanja obrazaca i problema koji se razlikuju u odnosu na sintetičke podatke. Uz pretpostavku postojanja veće količine računarskih resursa predlaže se primena prikazane metodologije nad neredukovanim podacima, naročito nad podacima o zapažanjima i merenjima. Nadalje, u slučaju korišćenja metoda klasifikacije gde je to moguće, predlaže se automatska optimizacija parametara modela, kao na primer u slučaju XGBoost modela.

LITERATURA

[1] Ruan, T., Lei, L., Zhou, Y. et al. "Representation learning for clinical time series prediction tasks in electronic health records". BMC Med Inform Decis Mak 19, 259 (2019).

[2] Chungsoo Kim, Seng Chan You, Jenna M Reps, Jae Youn Cheong, Rae Woong Park, "Machine-learning model to predict the cause of death using a stacking ensemble method for observational data", Journal of the American Medical Informatics Association (2020).

[3] Weng SF, Vaz L, Qureshi N, Kai J (2019) "Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches".

[4] Walonoski, Jason, et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic healthcare record." Journal of the American Medical Informatics Association 25.3 (2018)

[5] Seely, Cara. "Validation of Synthea." (2017)

[6] Chen, J., Chun, D., Patel, M. et al. "The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures.", BMC Med Inform Decis Mak (2019)

[7] Ng, Andrew. "Sparse autoencoder". CS294A Lecture notes (2011)