Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.9
DOI: https://doi.org/10.24867/30BE31Aleksic

PRIMENA MAŠINSKOG UČENJA U PREDIKCIJI ŽIVOTNOG VEKA LJUDI NA OSNOVU SOCIO-EKONOMSKIH I DEMOGRAFSKIH KARAKTERISTIKA

APPLICATION OF MACHINE LEARNING IN PREDICTING HUMAN LIFE EXPECTANCY BASED ON SOCIO-ECONOMIC AND DEMOGRAPHIC CHARACTERISTICS

Lenka Isidora Aleksić, Fakultet tehničkih nauka, Novi Sad

Oblast – PRIMENJENE RAČUNARSKE NAUKE I INFORMATIKA

Kratak sadržaj – *Ovaj rad analizira primenu različitih metoda mašinskog učenja u predikciji životnog veka ljudi koristeći socio-ekonomske i demografske karakteristike. Kroz obradu podataka iz skupa ‚World Health Statistics 2020' Svetske zdravstvene organizacije, razvijeni su modeli koji koriste algoritme poput Decision Tree, XGBoost, Random Forest i neuronskih mreža. Rad opisuje proces preprocesiranja podataka, treniranje modela i evaluaciju performansi, sa posebnim fokusom na upotrebu Python biblioteka kao što su NumPy, Pandas, TensorFlow i Scikit-learn za implementaciju rešenja.*

Ključne reči: *HALE; polu-nadgledano učenje; metode mašinskog učenja; životni vek; baza podataka SZO.*

Abstract – *This paper analyzes the application of various machine learning methods in predicting human life expectancy using socio-economic and demographic characteristics. Through data processing from the 'World Health Statistics 2020' dataset, models have been developed using algorithms such as Decision Tree, XGBoost, Random Forest, and neural networks. The paper describes the data preprocessing process, model training, and performance evaluation, with a particular focus on the use of Python libraries such as NumPy, Pandas, TensorFlow, and Scikit-learn for implementing the solutions.*

Keywords: *HALE; semi-supervised learning; machine learning methods; life-expectancy; WHO-dataset;*

1. UVOD

Tema životnog veka je od važna za javnozdravstvene ustanove, jer omogućava bolje razumevanje faktora koji utiču na dugovečnost ljudi, a time i razvoj politika koje mogu poboljšati zdravstvene ishode u različitim populacijama. Za potrebe ovog istraživanja korišćeni su podaci Svetske zdravstvene organizacije iz skupa World Health Statistics 2020 (WHS2020), koji obuhvata informacije za preko 200 zemalja sveta. Cilj rada je primeniti različite modele mašinskog učenja radi predikcije očekivanog životnog veka na osnovu gorenavedenih karakteristika. Poseban akcenat stavljen je na optimizaciju modela i njihovu sposobnost predikcije na globalnom nivou. Metode poput *Decision Tree*-a, *XGBoost*-a, i dubokih neuronskih mreža primenjene su kako bi se omogućilo otkrivanje skrivenih obrazaca u podacima. Rad se takođe osvrće na izazove rada sa podacima, kao što su nepotpune vrednosti i normalizacija. Ovo istraživanje nastoji da ponudi dublji uvid u predikciju životnog veka korišćenjem savremenih tehnologija mašinskog učenja, što može imati značajan uticaj na različite discipline, uključujući ekonomiju, sociologiju i javne politike.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Dunja Vrbaški, asis. prof.

2. SKUP PODATAKA

Skup podataka za implementaciju rešenja sadrži 39 zasebih fajlova o različitim temama koji mogu uticati na zdravlje ljudi. Glavni fokus podataka je na indikatorima kao što su očekivani životni vek, stope mortaliteta, zdravstvene usluge, i socio-ekonomski faktori. Svaka tabela unutar skupa podataka sadrži informacije o zemljama, godinama, demografskim promenljivim, kao i o zdravstvenim pokazateljima, omogućavajući sveobuhvatnu analizu.

Health Adjusted Life Expectancy (HALE) kombinuje dužinu života sa kvalitetom života, pružajući dublji uvid u zdravlje stanovništva i predstavlja glavni indikator za metode mašinskog učenja, za koji je ovaj skup podataka posebno prilagođen. Time se omogućava analiza velikih i složenih zdravstvenih trendova putem regresije, klasifikacije i klasterizacije. Međutim, jedan od najvećih izazova pri radu sa ovim podacima jeste prisustvo nedostajućih vrednosti i neujednačenost u formatima, što zahteva pažljivo pretprocesiranje, uključujući normalizaciju i imputaciju nedostajućih vrednosti. Uprkos izazovima, primena ovih tehnika omogućava stvaranje robusnog skupa podataka koji može služiti za precizniju analizu zdravstvenih trendova i predikciju životnog veka na globalnom nivou.

3. KORIŠĆENE TEHNOLOGIJE ZA ANALIZU PODATAKA

Prilikom rada na analizi i modeliranju podataka u okviru predikcije životnog veka, *Python* biblioteke predstavljaju alat za efikasno upravljanje podacima, statističku analizu i primenu mašinskog učenja. One omogućavaju obradu velikih skupova podataka, vršenje kompleksnih matematičkih operacija i vizualizaciju rezultata, što je neophodno za savremeni pristup analizi zdravstvenih

podataka. Kako zajedno čine jedinstvenu osnovu za obradu i analizu podataka u okviru ovog projekta, u sledećim poglavljima biće opisane njihove ključne funkcije.

3.1. *NumPy* biblioteka

NumPy je osnovna *Python* biblioteka koja omogućava efikasnu manipulaciju višedimenzionalnim nizovima podataka. Njena ključna struktura je *ndarray*, koja ubrzava operacije na velikim setovima podataka zbog mogućnosti skladištenja podataka u kontinuiranom bloku memorije. Ova organizacija omogućava brzo izvršavanje složenih matematičkih operacija poput matričnih operacija, Fourierovih transformacija, kao i rešavanje sistema linearnih jednačina. Jedna od najvažnijih karakteristika *NumPy*-a je vektorizacija, koja omogućava operacije na celim nizovima podataka bez potrebe za eksplicitnim iteracijama. To čini rad sa velikim podacima efikasnim i ubrzava analizu, posebno u aplikacijama mašinskog učenja. U ovom radu, *NumPy* je bio presudan za pripremu podataka i statističku analizu, osiguravajući brzinu i efikasnost u procesu obrade velikih numeričkih skupova podataka.

3.2. *Pandas* biblioteka

Pandas služi za rad sa tabelarnim podacima, posebno korisna za manipulaciju i analizu strukturiranih podataka kroz strukturu *DataFrame*. Ova struktura omogućava skladištenje podataka različitih tipova u redovima i kolonama, omogućavajući efikasno baratanje velikim datasetovima. Pandas nudi funkcije za rad sa nedostajućim vrijednostima, poput *fillna()* i *dropna()*, što je od velike važnosti u radu sa stvarnim podacima. Biblioteka takođe pruža alate za grupisanje i agregaciju, olakšavajući složene analize po kategorijama. U ovom radu, *Pandas* je bio ključan u pripremi podataka za modele mašinskog učenja, omogućavajući spajanje i transformaciju različitih datasetova, te analiziranje vremenskih serija koje su bile za istraživanje demografskih i zdravstvenih trendova.

3.3. *Scikit-Learn* biblioteka

Scikit-learn je osnovna *Python* biblioteka za mašinsko učenje. Veoma je korisna jer nudi širok spektar algoritama za klasifikaciju, regresiju, klasterizaciju, kao i alate za smanjenje dimenzionalnosti. Posebno se ističe zbog svoje jednostavnosti upotrebe i standardizovanog procesa treniranja modela.

U ovom radu, *Scikit-learn* je korišćen za implementaciju različitih modela mašinskog učenja poput *Decision Tree*, *Random Forest*, i *XGBoost*. Uz pomoć funkcija kao što su *train_test_split()* i *cross_val_score()*, proces treniranja i evaluacije modela je bio olakšan, omogućavajući lako poređenje performansi. Biblioteka je takođe omogućila napredne tehnike za pretprocesiranje podataka, skaliranje i enkodiranje kategorijalnih promenljivih. Evaluacija modela je vršena uz metrike poput *Mean Squared Error* (MSE) i R^2 *Score*.

3.4. *TensorFlow* i *Keras* biblioteke

TensorFlow i *Keras* su korišćeni za izradu i treniranje dubokih neuronskih mreža, omogućavajući fleksibilan rad sa složenim podacima. *Keras* je API koji olakšava kreiranje i treniranje modela, dok *TensorFlow* pruža infrastrukturnu podršku i ubrzanje uz pomoć GPU-a. Ove biblioteke omogućavaju dodavanje slojeva i podešavanje hiperparametara modela, što je bilo presudno u optimizaciji performansi neuronskih mreža korišćenih u ovom radu. Funkcije poput *fit()* i *evaluate()* olakšale su proces treniranja, dok su napredne tehnike, kao što su *Recurrent Neural Networks* (RNNs) i *Convolutional Neural Networks* (CNNs), korišćene za analizu vremenskih serija i prepoznavanje obrazaca u podacima.

4. METODE OBUČAVANJA

U ovom poglavlju akcenat će biti na modelima obučavanja koji su korišćeni u ovom projektu kako bi se dobili optimalni rezultati. Sposobnost mašinskog učenja da uči iz podataka i donosi odluke na osnovu tih znanja čini ga jednim od najmoćnijih alata u današnjem svetu informacionih tehnologija.

4.1. *Decision Tree* model obučavanja

Decision Tree je osnovni algoritam mašinskog učenja koji se koristi za rešavanje problema klasifikacije i regresije. Ovaj model oponaša način na koji ljudi donose odluke kroz hijerarhijsku strukturu drveća, gde svaka grana predstavlja odluku na osnovu ulaznih podataka, a svaki čvor predstavlja karakteristiku ili atribut. *Decision Tree* je jednostavan za interpretaciju i vizualizaciju, zbog čega se često koristi u oblastima kao što su medicina i finansije, gde je objašnjivost modela ključna. Ipak, jedan od njegovih najvećih nedostataka je sklonost ka prekomernom učenju (*overfitting*), naročito kada model postane previše složen ili dubok. Da bi se smanjila ova sklonost, koriste se tehnike poput orezivanja stabla (*pruning*) ili ograničenja na maksimalnu dubinu stabla.

4.2. Neuronske mreže

Neuronske mreže su napredan alat u mašinskom učenju, inspirisan radom bioloških neurona, i omogućavaju rešavanje složenih zadataka poput prepoznavanja obrazaca, obrade slika i prirodnog jezika. Ove mreže se sastoje od slojeva veštačkih neurona koji su povezani u mrežu i mogu učiti složene i nelinearne relacije u podacima. U ovom radu, neuronske mreže su korišćene za predikciju životnog veka na osnovu demografskih i zdravstvenih podataka. Zbog svoje sposobnosti da modeliraju složene odnose, neuronske mreže su postale ključne u mnogim oblastima, ali njihova primena zahteva veliku količinu podataka i vremena za treniranje.

4.3. XGBoost u mašinskom učenju

XGBoost je napredna implementacija algoritma gradijentnog boostinga, koja je poznata po visokoj brzini i preciznosti. Ovde *XGBoost* je korišćen je kao još jedan model jer pruža veoma precizne rezultate. Ključni princip *XGBoost*-a je kombinovanje više jednostavnih modela, poput *Decision Tree*-a, kako bi se formirao snažan ansambl. Algoritam koristi različite tehnike optimizacije, poput regulacije (L1 i L2), kako bi se izbegao problem pretreniranosti i poboljšale performanse modela.

4.4. *Random Forest* model obučavanja

Random Forest je ansambl metoda koja kombinuje više *Decision Tree*-a kako bi se poboljšala stabilnost i tačnost

predikcija. Prednost ovog modela je u tome što primenjuje tehniku *bagging*, koja koristi različite podskupove podataka za generisanje svakog stabla, čime se smanjuje rizik od pretreniranosti. Konačna predikcija se zasniva na glasanju svih stabala u ansamblu, čime se postiže visoka tačnost i otpornost na šum u podacima.

4.5. ARIMA model u analizi vremenskih serija

ARIMA (*Autoregressive Integrated Moving Average*) model je jedan od najpoznatijih modela za analizu vremenskih serija. Ovaj model se koristi za predikciju vremenski zavisnih podataka, kao što su ekonomski indikatori ili vremenske promene. Sastoji se od tri glavna dela: autoregresivnog (AR), integrisanog (I) i pokretnog proseka (MA), gde svaka komponenta modeluje određene aspekte vremenskih serija. ARIMA model je korišćen za predikciju životnog veka jer omogućava hvatanje obrazaca u podacima i preciznu prognozu budućih vrednosti na osnovu istorijskih podataka.

5. REZULTATI I DISKUSIJA

Gore navedeni modeli testirani su pomoću nadgledanog i polu-nagledanog učenja, s ciljem da se identifikuju najbolje performanse i ključni faktori uticaja na predikcije.

Jedan od ključnih aspekata analize bio je procena uticaja vremenske kolone, koja predstavlja period u kojem su prikupljeni podaci, na performanse modela. RMSE i R^2 metrike su korišćene za evaluaciju tačnosti modela. Na osnovu rezultata, modeli kao što su *Random Forest* i *XGBoost* pokazali su se superiornima u poređenju sa *Decision Tree* modelom i potpuno povezanom neuronskom mrežom. Takođe, polu-nadgledano učenje nije značajno poboljšalo performanse u većini slučajeva. Samo kod *Decision Tree* modela i neuronske mreže bez period kolone uočeni su pozitivni rezultati polu-nadgledanog učenja. Kada je period kolona uključena, neuronske mreže su pokazale poboljšanja u predikciji. Najbolje performanse su primećene kod *XGBoost* i *Random Forest* modela, sa stabilnim rezultatima u različitim varijacijama podataka.

Rezultati RMSE i R^2 metrika pokazali su konzistentnost predikcija kod modela sa period kolonom, dok su modeli bez ove kolone pokazali manju tačnost. Pored toga, rezultati su pokazali da je vremenska komponenta značajan faktor u predikciji očekivanog životnog veka, posebno u modelima koji uključuju vremenske serije.

Generalno, performanse modela su bile osetljive na kvalitet podataka, sa boljim rezultatima kada je dostupna veća količina podataka. Primena naprednih tehnika kao što su vektorizacija i regulacija parametara, značajno su doprinele poboljšanju preciznosti predikcija.

6. ZAKLJUČAK

U radu su istražene metode mašinskog učenja i veštačke inteligencije za predikciju očekivanog životnog veka na osnovu podataka iz World Health Statistics 2020. Kroz primenu modela kao što su Decision Tree, XGBoost, Random Forest, i potpuno povezana neuronska mreža, analizirani su socio-ekonomski i demografski faktori koji utiču na HALE metriku. Rezultati pokazuju da faktori poput obrazovanja, prihoda i zdravstvenog stanja značajno utiču na očekivani životni vek, dok su pol i mesto stanovanja takođe važni prediktori. Primena semi-supervised metoda donela je dodatne uvide, ali postoji mogućnost za dalje unapređenje modela uključivanjem novih podataka, kao što su zdravstvene navike i genetske predispozicije. Optimizacija hiperparametara i napredne tehnike obrade podataka mogu doprineti boljoj generalizaciji modela. Zaključno, ovaj rad naglašava potencijal mašinskog učenja u predikciji zdravstvenih ishoda i pruža osnovu za buduće alate koji će unaprediti donošenje odluka u zdravstvu i javnim politikama.

7. LITERATURA

[1] Wolfson, M.C., 1996. Health-adjusted life expectancy. Health Reports-Statistics Canada, 8, pp.41-45.

[2] Luy, M., Di Giulio, P., Di Lego, V., Lazarevič, P. and Sauerberg, M., 2020. Life expectancy: frequently used, but hardly understood. Gerontology, 66(1), pp.95-104.

[3] Tanha, J., Van Someren, M. and Afsarmanesh, H., 2017. Semi-supervised self-training for decision tree classifiers. International Journal of Machine Learning and Cybernetics, 8, pp.355 370.

[4] Hill, K. and Choi, Y., 2006. Neonatal mortality in the developing world.Demographic research, 14, pp.429-452.

Kratka biografija:

Lenka Isidora Aleksić rođena je 14.10.1999. godine u Zrenjaninu. Smer računarstvo i automatika na Fakultetu tehničkih nauka u Novom Sadu upisala je 2018. godine. Osnovne studije završila je u septembru 2022. godine. Od oktobra 2022, nakon upisa na master studije, kreće redovno da radi na fakultetu kao saradnik u nastavi, a ubrzo i u struci kao *Data* inženjer.