Предикција животног века људи на основу различитих социо-економских и демографских карактеристика

Иван Мршуља (Аутор)
Факултет Техничких Наука
Катедра за информатику
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21000, Нови Сад, Војводина, Србија
ivan.mrsulja@uns.ac.rs

Ленка Исидора Алексић (Аутор)
Факултет Техничких Наука
Катедра за примењене рачунарске науке
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21000, Нови Сад, Војводина, Србија
lenkaisidora.aleksic@uns.ac.rs

Милош Поповић (Аутор)
Факултет Техничких Наука
Катедра за информатику
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21000, Нови Сад, Војводина, Србија
milospopovic@uns.ac.rs

Апстракт—Предикција животног века је од велике значајности у области јавног здравља, јер нам омогућава да боље разумемо и предвидимо здравствено стање становништва. У овом пројекту коришћен је скуп података World Health Statistics 2020, а примењене су различите технике машинског учења - Decision tree, XGBoost, Random Forest и потпуно повезану невронску мрежу - како би се регресионо предвидела HALE метрика на основу социо-економских и демографских обележја. Коришћен је полу-надгледан приступ приликом обучавања модела код два регресиона приступа: класична и time-series регресија. Циљ пројекта био је да се добије што прецизнија предикција HALE метрике, која се користи за мерење здравственог стања становништва. За евалуацију перформанси модела коришћена је RMSE метрика као и R2 вредност. Резултати су показали да су XGBoost и Random Forest дали најбоље перформансе у односу на остале технике (2.26 RMSE и 0.9 R2; 2.18 RMSE и 0.91 R2 респективно). Овај рад има за циљ допринети разумевању фактора који утичу на HALE метрику и може бити користан за даља истраживања у области јавног здравља.

Кључне речи—HALE; semi-supervised учење; life-expectancy; машинско учење; who-dataset; time-series

I. Увод

Предикција животног века људи је једна од кључних тема у истраживању демографије, јавног здравља и социологије [3]. Животни век људи је сложен феномен који зависи од бројних социо-економских и демографских фактора, као што су године образовања, приходи, статус запослености, друштвени положај, здравствено стање, пол, етничка припадност и место становања. У последњих неколико деценија, постигнут је значајан напредак у области предикције животног века уз помоћ напредних статистичких метода и метода машинског учења.

У овом раду, анализираћемо различите социо-економске и демографске карактеристике из *World Health Statistics* 2020 (у наставку *WHS2020*) [4] скупа података које утичу на животни век људи и применићемо неке од модела машинског учења (*Decision tree, XGBoost, Random Forest* и потпуно повезана неуронска мрежа) како бисмо предвидели очекивани животни век на основу тих карактеристика. С обзиром да је већина статистика из *WHO (World Health Organization)* скупа података вађена чешће од *HALE* метрике, у подацима се јавља проблем великог броја нелабелираних вредности што ћемо покушати да решимо применом полу-надгледаног (енг. *semi-supervised*) приступа приликом обучавања модела. Такође, хтели бисмо да узмемо у обзир и временску информацију о томе када је одређена статистика вађена, стога ћемо поред стандардног регресионог приступа укључити и *time-series* приступ где ћемо испитати да ли податак о временском раздобљу када су подаци прикупљени има утицаја на побољшање квалитета предикције различитих модела. На тај начин, циљ нам је да допринесемо бољем разумевању и предвиђању животног века људи, што би могло бити корисно у различитим областима, укључујући јавно здравље, економију и социологију.

Битна ствар је нагласити да се "животни век" у овом раду односи на *HALE (Health Adjusted Life Expectancy)*, а не на стандардну *LE (Life Expectancy)* метрику. *LE* је метрика која се користи да се израчуна очекивани број година које особа може очекивати да живи, на основу старосне структуре становништва и стопе смртности у популацији [5]. Иста је заснована на претпоставци да ће стопа смртности остати иста током живота свих особа у популацији, што наравно није увек случај. Такође, *LE* не узима у обзир здравствено стање особа, нити друге факторе који могу утицати на квалитет живота. Са друге стране, *HALE* је метрика која узима у обзир не само дужину живота, већ и квалитет живота и израчунава се на основу просечне дужине живота и очекиваног здравственог стања особе у одређеном добу [3]. Оваква метрика пружа бољи увид у стварно здравствено стање

популације и узима у обзир и инвалидитет, болести и друге факторе који утичу на квалитет живота. *HALE* се користи као мера здравственог стања популације и као индикатор ефикасности здравственог система у земљи [3].

У поглављу II дат је преглед постојеће релевантне литературе која нам је помогла да приступимо проблему и дефинишемо релевантну методологију. Након тога у III поглављу ћемо представити методологију коју смо користили за решавање проблема. У наредном поглављу IV ће детаљно бити описан скуп података а методе евалуације модела у поглављу V. У наредном поглављу VI представићемо резултате различитих коришћених модела машинског учења и дати коментар и детељнију елаборацију на перформансу сваког од њих. На крају, поглавље VII закључује овај рад.

II. Преглед постојеће релевантне литературе

С обзиром да су се сви радови који су се бавили *WHS2020* скупом података, били експлоративна анализа, покушали смо наћи радове који су на сличним скуповима података покушали предвидјети *LE* ili *HALE* метрику. У наставку биће представљена два најрелевантнија рада из прегледане групе. Такође, биће представљен рад који објашњава притуп полу-надгледаног учења на *time-series* подацима, одакле смо и добили мотивацију да пробамо исти.

A. *Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth. [1]*

Тема овог рада је методологија коришћења дубоких неуронских мрежа (DNNs) за процену смртности на основу очекиване дужине живота. Смртност се односи на вероватноћу смрти у сваком појединачном узрасту, а не на укупну очекивану дужину живота, која представља просек свих стопа смртности. Аутори овог рада предлажу да се њихов модел може користити за процену смртности за земље или популације које немају поуздане податке о овом показатељу, користећи информације из суседних популација или оних са сличном динамиком смртности.

Тестирана су три модела за решавање овог проблема:

- Потпуно повезана неуронска мрежа
- *Linear-Link*
- *Ševčíková* et al. (2016) коришћењем *Lee-Carter* модела

Аутори су користили податке из базе података о људској смртности (*Human Mortality Database - HMD 2021*). База података садржи ифнормације о подацима из Јапана, Италије, Америке и Русије, подаци су категорисани по полу. Улазак у неуронску мрежу је матрица стопе смртности где редови представљају број година а колоне календарску годину.

За евалуацију је коришћен *MAE* (*Mean Absolute Error*) и *RMSE* (*Root Mean Squared Error*).

Поређени су резултати потпуно повезане неурноске мрежа са друга два модела (*Linear-Link* и *Ševčíková* model). Резултати су углачани (*smoothened*) коришћењем *P-splines* како би се обезбедила упоредивост.

Овај рад, иако се не бави потпуно истим проблемом као наш, довољно је сличан у контексту формата података и као такав нам је помогао да увидимо комплексност проблема као и могуће методе евалуације које можемо корстити. Такође, био је мотивација да користимо потпуно повезану неуронску мрежу као један од модела у нашем приступу.

B. *Health-adjusted life expectancy (HALE) in Chongqing, China* [2]

Аутори овог рада развијали су тему израчунавања очекиваног животног века популације кинеског града *Chongqing* са уделом здравља саме популације (*HALE*). Такође, поред *HALE* рачунате су и *DLE* (*Disability Life Expectancy*) и *LE*. Новијим приступом и укомбиновањем података везаних за здравље и смртност, добили су значајне резултате који могу потпомоћи и пружити увид у старење становништа и здравствене проблеме на глобалном нивоу.

Користећи информације о популацији, аутори су имплементирали методу екстракције болести користећи технике *NLP*-а (*Natural Language Processing*). Како су подаци из медицинских докумената неструктуирани, одрађена је нормализација и након тога издвајали су се подаци од интереса од слободног текста коришћењем *RNN* (*Recurrent Neural Network*). Након добијања имена болести, оне су мапиране на таксономске кодове болести уз додатну помоћ доменских стручњака и након овога омогућена је екстракција распрострањености саме болести. Укомбиновањем података о болестима са статистичким подацима о смрти, дали су могућност израчунавања вероватноће смртности свих доба становништва са утицајем здравља (односно болести) на скраћење животног века људи.

Потребни подаци за израчунавање *HALE* метрике у овој студији добијени су из *FIS* (*Family Population Information System*) система, док су *EMR* (*Electronic Medical Records*) подаци одобрени за употребу од стране здравствене комисије *Chongqing* округа након што су личне информације пацијената анонимизоване (важи што су ★). Такође, коришћени су и подаци о људској смртности из различитих база података (*Human Mortality Database-HMD, FIS* и др.) као и *WHO* база података о инвалидитетима.

Закључак студије добро се слаже са стварним статистичким подацима, док су резултати Спирмановог теста коефицијента корелације ранга, статуса здравља

екстрахованог из ове студије и здравља популације *Chongqing* града, у потпуности статистички у корелацији.

Иако су подаци уско везани за једну област у Кини, обележја су јако слична нашима па смо из овог рада добили важне информације о *feature-engineering*-у самих података као и начине којима бисмо спречили увођење *gender-bias*-а у наше моделе.

C. *Semi-Supervised Time Series Classification by Temporal Relation Prediction [7]*

Тема рада је полу-надгледана *time-series* класификација, тј. приступ машинског учења који комбинује мало означених података са великим бројем неозначених података током тренирања у циљу побољшања перформанси модела. Аутори предлажу методу названу *SemiTime*, која користи полу-надгледани приступ и структуру неозначених података у циљу побољшања перформанси модела који је трениран методама надгледаног учења.

Скуп података је подељен на лабелирани и нелабелирани део. Тест скуп је добијен семпловањем једног дела скупа лабелираних података док се тренинг обављао над остатком лабелираних података и свим нелабелираним подацима. Модел је подељен на 3 модула и то:

- *temporal relational segment sampling*
- Надгледано обучавање
- Ненадгледано обучавање

Најпре се над лабелираним делом тренинг скупа спроведе обучавање произвољним моделом. Аутори рада су користили неименован модел и само су нагласили да се састоји од feature extractor-a и главе за класификацију. Након иницијалног надгледаног учења врши се иницијала евалуација над тестим скупом. Затим се користи самогенерисана веза временског сегмента (*temporal relational segment sampling*) као надзорни сигнал и спроводи се задатак предвиђања временске релације над неозначеним временским серијама. Сваки пут када се временска серија ti подели на 2 дела (где предњи део B-дужине означава прошли сегмент, а задњи део (T − B)-дужине означава будући сегмент) где је B = [α ∗ T] (α је однос прошлих и будућих сегмената) израчунава се веза временског сегмента тако што се за сваки део селектује будући парњак s+i,α (из исте временске серије) и прошли парњак s-i,α (из различите временске серије) у односу на anchor сегмент si,α по формули: zi,α = fθ(si,α), z+i,α = fθ(s+i,α), z−j,α = fθ(s−i,α) и касније се користи као додатно обележје.

Затим се врши поновна екстракција обележја и класификација нелабелираних делова углавном за класификацију уз додатну бинарну класификацију која говори да ли је релација класификована као позитивна или негативна у конкретној временској серији којој припада. На крају, тренирање је поновљено над целим скупом података и вршена је коначна евалуација након чега се нови резултати упоређују с иницијално добијеним.

Експерименти су вршени на *CricketX*, *UWaveGestureLibraryAll*, *InsectWingbeatSound*, *XJTU*, *MFPT* и *EpilepticSeizure* скуповима података.

Као мера перформансе коришћена је тачност гдје су се остварили резултати од 65.66%, 90.29%, 66.57%, 98.46%, 84.33%, 79.26% на сваком од скупова респективно што је у просеку било ~7% боље од иницијалног приступа са надгледаим учењем.

III. Методологија

С обзиром да се скуп података састоји из 39 раздвојених фајлова који осликавају различите социо-економске карактеристике у државама у периоду од 2000 - 2019 године, идеја је да се помоћу имена државе и године за коју је податак извучен креира коначни скуп података где се као циљно обележје користи *HALE* обележје. С обзиром да нису све статистике ваљене редовно, јавља се проблем недостајућих вредности где смо испитали различите технике апроксимације истих. Такође, одрађен је својеврсни *data-engineering* гдје смо уклонили неке од колона које нису биле информативне или су биле јако ретко/давно ваљене да нису биле релевантне. Лабелирани подскуп је издељен на тренинг, валидациони и тест скуп, а спајањем тренинг података и нелабелираних података се добија коначан тренинг скуп са којим се обучавају наши модели применом *self-training* технике полу-надгледаног учења [6]. Модели које смо обучавали су: стабло одлучивања (основни модел за иницијалну процену комплексности проблема), *Random Forest*, *XGBoost* і потпуно повезана неуронска мрежа (*Fully Connected Neural Network*). Такође, с обзиром да имамо временску компоненту коју у претходним приступима не користимо, идеја је била да пробамо полу-надгледани приступ *time-series* анализе сличан приступу [7]. С обзиром да се у претходно поменутом раду решава класификациони проблем, у нашем пројекту тестирана је модификована верзија где смо користили *Exponential Smoothing* (*ETS*) и *AutoARIMA* модел, који узимају у обзир *temporal relational segment sampling*. Имплементација је одрађена али нисмо успели ефикасно да обучимо модел а ни да га тестирамо из разлога што нам за *time-series prediction* приступ треба такав скуп података, где имамо X сукцесивних тренинг инстанци мерених у приближно истом временском размаку и исто тако одређен број тестних инстанци, мерених у истом или сличном размаку. Како је наш скуп података прилично неконзистентан због саме природе проблема (у великом броју држава се због мањка финансијских средстава не ваде све *WHO* предложене метрике сваке године [4]), одлучено је да се пређе на *time-series regression* приступ, где је већ постојећи тренинг скуп података проширен енкодованом временском компонентом а затим су исти

регресиони модели од раније били поново тренирани и оптимизовани где смо забележили који модели су имали користи од увођења новог обележја, а који нису. Иницијално тренирање је обављено над лабелираним подацима где је извршена иницијална евалуација над тестним скупом. Након тога, помоћу истренираног модела врши се предвиђање лабела нелабелираног скупа. На крају, тренирање се понавља над новим, проширеним, тренинг скупом и ради се коначна евалуација која нам показује да ли је наше полу-надгледано учење имало утицаја на побољшање перформанси иницијалног модела. Битно је напоменути да се у фази оптимизације хипер параметара, евалуација одвијала над валидационим скупом, док су у финалној верзији пројекта тренинг и валидациони скуп спојени у већи тренинг скуп а за евалуацију користи се тестни скуп како би се што реалније приказале перформансе модела. Битно је нагласити да је у фази оптимизације испитана могућност више итерација полу-надгледаног тренинга, међутим, након прве итерације није забележено побољшање те је овај приступ одбачен.

Са софтверске стране, проблем је решен коришћењем *Python3* програмског језика уз ослонац на библиотеке за обраду података и машинско учење: *NumPy*, *ScikitLearn*, *Pandas*, *XGBoost* и *StatsForecast*.

IV. Опис скупова података

Иницијални скуп података обухвата најновије и ажуриране здравствене карактеристике света (признатих држава стране СЗО). Сами подаци филтрирани су на основу различитих показатеља и подељени су на подкатегорије. Коришћењем експлоративне анализе уочено је доста недостатака самог скупа података, те смо на почетку рада одмах одбацили одређене фајлове који нису били релевантни или су имали *null(NaN)* вредности које су преовладавале. Такође постојали су фајлови у којима мерења нису редовно извршавана док у другим случајевима велики број држава није ни имао информације о тестираном параметру. Након филтрације фајлова остало их је 20 који заправо дају некакву вредност при израчунавању *HALE* метрике. Неки од примера фајлова су: *basicDrinkingWaterServices.csv* (проценат становништва које користи барем основне улоге воде за пиће), *birthAttendedBySkilledPersonel.csv* (проценат порођаја на којима је присуствовало стручно особље), *infantMortilityRate.csv* (вероватноћа смрти у првој години живота на 1000 рођења), *tobaccoAge15.csv* (преваленција тренутне употребе дувана међу особама старијим од 15 година) итд.

Након издвајања сетова података од интереса, извршено је њихово спајање у један *dataframe* по *Location* и *Period* колонама.

Како су подаци пре 2000. године били ретки и са пуно недостатака, над новонасталим фајлом, извршено је филтрирање тако да су узете само године након 2000. године. Такође, информације о проценту деловања одређене карактеристике, биле су подељене по мушким и женским особама док су постојали подаци и у укупном броју невезано за пол особе. Ради лакшег каснијег анализирања, користили смо само спојене податке, а исти метод смо применили и на податке из градова и села.

Интерполацију смо користили како бисмо попунили недостајеће вредности и њу смо применили на све податке осим на саму *HALE* метрику за коју се врши предикција. Поред интерполације извршена је и нормализација скупа података.

V. Методе евалуације

MAE (Mean Absolute Error) је метрика евалуације која се користи за оцењивање перформанси предиктивних модела. Мери апсолутну разлику између предвиђених вредности и стварних вредности и даје једну вредност (тачност предвиђања). Резултат *MAE* метрике се евалуира на основу његове вредности. Што је вредност *MAE* ближа нули, то је предиктивни модел тачнији. Када се користи *MAE* за оцењивање модела, прво се израчуна *MAE* за сваку инстанцу података у тест скупу. Затим се израчуна просечна вредност *MAE* за цео тест скуп. Та просечна вредност представља квалитет модела. Критеријуми за евалуацију *MAE* могу да се разликују у зависности од контекста примене модела.

RMSE (*Root Mean Squared Error* - квадратни корен из средњеквадратне грешке) је метрика која мери разлику између предвиђених вредности и стварних вредности и даје једну вредност која представља укупну тачност предвиђања модела. Често се користи у развоју модела машинског учења, као што су линеарна регресија, неуронске мреже и дрво одлуке. Резултат *RMSE* метрике се евалуира на основу његове вредности. Исто као и код *MAE* што је вредност *RMSE* ближа нули, то је предиктивни модел тачнији. Такође, прво се израчуна *RMSE* за сваку инстанцу података у тест скупу и затим се израчуна просечна вредност *RMSE* за цео тест скуп. Видимо да је *RMSE* је сличан *MAE*, али се разликује у начину на који се израчунава. *RMSE* прво квадрира грешку за сваку инстанцу, затим израчунава средњу вредност тих квадратних грешака, и на крају израчунава корен из те вредности. Овај процес додатно поглашава велике грешке, што *RMSE* чини метриком која је осетљивија на велике грешке него *MAE*.

R2 (R-squared, још се назива и коефицијент детерминације*)* метрика је мера колико добро модел прилагођава податке. *R2* метрика се користи за оцењивање регресионих модела, тј. модела који предвиђају непрекидне вредности. *R2* метрика је однос између објашњене варијансе и укупне варијансе података. Објашњена варијанса представља количину варијансе узроковане независним променљивама, која је објашњена моделом. Укупна варијанса представља количину

варијансе у подацима која се може објаснити моделом. Вредност *R2* метрике се креће између 0 и 1. Што је вредност *R2* метрике ближа 1, то је модел тачнији. Ако је *R2* метрика 1, то значи да је модел савршено прилагођен подацима. Евалуација *R2* метрике се врши упоређивањем *R2* метрике добијене на тренинг скупу података са *R2* метриком добијеном на тест скупу података. Циљ је да се добију сличне вредности *R2* метрике на оба скупа података, што је индикација да је модел добро генерализован и да ће добро предвиђати на новим, непознатим подацима. *R2* метрика је корисна зато што је једноставна за тумачење и лако је мерљива. Међутим, треба имати на уму да *R2* метрика није увек најбољи избор за оцењивање модела. На пример, уколико имамо неравномерно распоређене податке, *R2* метрика може да даје лажно позитивне или негативне резултате.

Када се упоређују два или више модела, боље је да се користи иста метрика, како би се добили резултати који су упоредиви. Уколико се модели оцењују користећи различите метрике, потребно је да се размотре њихове предности и мане у односу на конкретну примену, и да се изабере метрика која најбоље одговара датом проблему.

VI. Резултати и дискусија

Уз помоћ библиотека scikit-learn[1] коришћени су модели “Decesion Tree”, “Neural Network”, “Random Forest” і “MLP Regressor” за неуренску мрежу, поред тога коришћена је библиотека XGBoost[2] за XGBRegressor.
Резулатти су подељени на неколико варијација. Тестирали смо стандардно и полу надгледано учење за сваки модел. Поред тога смо тестирали значај колоне која представља период у ком су подаци настали, због ситуације где подаци нису били уједначени односно постоје периоди од неколико година без података.

У наредним табелама је приказана комбинација стандардног и полу надгледаног учења и евалуација са и без колоне која означава период података. Приказано је у RMSE и R2 метрици. Међутим изостављене су табеле са MSE мером пошто теоретски представљају само квадрирану RMSE меру, те нам не значи додатно за поређење резултата.
Звездицом је означен бољи резултат у односу на стандардно и полу надгледано учење.

Резултати (RMSE мера, без период колоне)

Алгоритам	Резултати	
	Стандардно	***Полу надгледано***
Decision Tree	2.974886952634062*	3.369713950596332
Neural Network	2.6721873253717123	2.663597699906164*
XG Boost	2.2618149166449144*	2.2658366593182304
Random Forest	2.212016440699879*	2.27675392900605

Мање је боље

Резултати (RMSE мера, са период колоном)

Алгоритам	Резултати	
	Стандардно	***Полу надгледано***
Decision Tree	3.116169810153851	2.9289733754361107*
Neural Network	2.44162634747905*	2.8115757026886015
XG Boost	2.247777416646409*	2.272524356603319
Random Forest	2.1866557063558973*	2.245555019075308

Мање је боље

Резултати (R2 мера, без период колоне)

Алгоритам	Резултати	
	Стандардно	***Полу надгледано***
Decision Tree	0.8368442876327051*	0.7906623127920419
Neural Network	0.8683577946110148	0.8692027502359393*
XG Boost	0.9056861491065569*	0.9053504512040756
Random Forest	0.909793454644219*	0.904436171697093

Веће је боље (од 0 до 1)

Резултати (R2 мера, са период колоном)

Алгоритам	Резултати	
	Стандардно	***Полу надгледано***
Decision Tree	0.8209791637807645	0.8418416236844216*
Neural Network	0.8900944167278289*	0.8542659906965705
XG Boost	0.9068531962571698*	0.9047909036563857
Random Forest	0.9118500306661226*	0.9070372953525034

Веће је боље (од 0 до 1)

На основу ових резултата може се закључити да су модели *XGBoost* и *random forest* остварили боље перформансе у односу на стабло одлучивања и потпуно повезану неуронску мрежу.

Такође прећијемо да полу надгледано учење није довело до бољих резултата у односу на надгледану учење

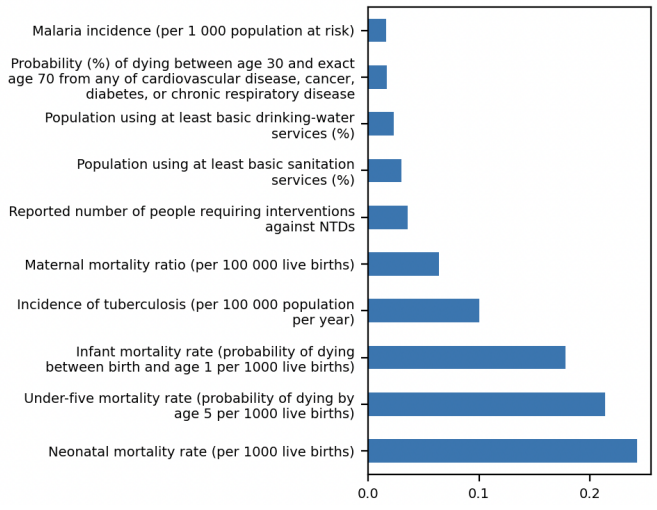
[1] Scikit-learn - је Python библиотека за машинско учење која садржи разне алгоритме за класификацију, регресију кластеровање и слично.
[2] Xgboost - је библиотека за изградњу модела за предвиђање, класификацију користећи алгоритам појачавања градијентом.

осим код стабла одлучивања, без период колоне, и код потпуно повезане неурноске мреже у случају када је укључена период колона. Међутим, потребно је узети у обзир друге факторе, као што су квалитет и количина доступних података, који су довелу до датих резултата.

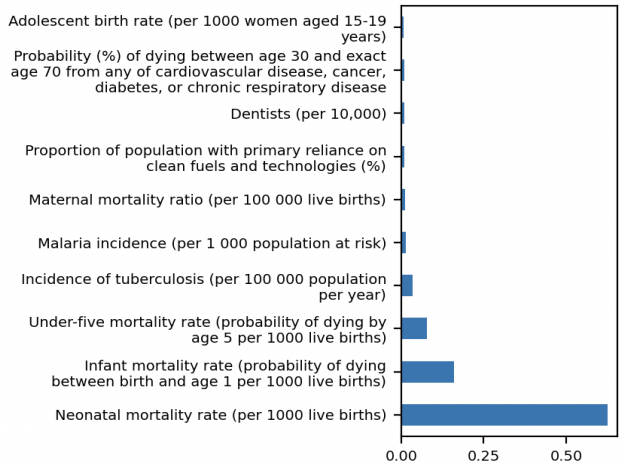
Коришћењем података који су нама били доступни највољи резултат је дао Random Forest алгоритам укључујући и колону са периодом прикупља где смо добили просечну грешку од 2.19 година за предикцију животног века на основу социо-економских фактора. Овај резултат, иако изгледа као јако добар, може се користити само као груба процена животног виека у областима где просечна грешка од две године не би представљала велики проблем, пример за то би биле области где се предикција обавља над великим бројем јединки нпр. област истраживања тржишта где се користи предвиђање животног века за одређивање потенцијалног тржишта за производе или услуге. У овом случају, иако би се грешка могла појавити у процени животног века циљне групе, она би вероватно имала минималан утицај на коначну одлуку о производњи или маркетингу.

У склопу NGBoost и осталих регресора из библиотеке *sklearn* постоји уграђена метода *feature_imporances* коју смо искористили за одређивање значајности обележја над подацима у обученом моделу. Ова метода користи *Bayes*-ову методу, која процењује вероватноћу губитка за сваку предикцију. *Bayes*-ова метода користи вероватноћу предикцији и вероватноћу параметара модела како би проценила вероватноћу губитка.

На сликама 1 и 2 су приказани 10 најзначајнијих обележја за модел *random forest* на слици 1 и *xgboost* на слици 2



Слика 1 - Најважнија обележја за "*random forest*" модел



Слика 2 - Најважнија обележја за *NGBoost* модел

Занимљиво је запазити да су оба модела највећи значај дали обележјима који осликавају смртност новорођенчади и дјеце до 5 година, што је један од главних показатеља колико је нека земља развијена како са медицинске тако и са економске стране [8]. Такође, треба напоменути да су модели највише грешили код земаља које представљају својеврсне "*outlier*-е" тј. веома богате/сиромашне земље, које доста одскачу од светског просјека а није опција да се уклоне из обучавајућег скупа јер у том случају губимо податке о истима. Могуће решење овог проблема јесте проширивање скупа података подацима за ове конкретне земље.

На крају, битно је нагласити да резултате нећемо дискутовати у контексту других радова из разлога што се не баве директно предикцијом HALE метрике.

VII. Закључак

У овом истраживању смо користили скуп података *World Health Statistics* 2020, који је објавила Светска здравствена организација (WHO) и садржи информације о здравственим показатељима и социо-економским факторима за различите земље. Скуп података садржи преко 50 различитих показатеља, укључујући факторе као што су очекивани животни век, стопа смртности од различитих болести, степен загађења воде, итд. У циљу анализе утицаја социо-економских и демографских фактора на *HALE* метрику, користили смо податке из различитих земаља широм света. У сврху предикције HALE метрике користили смо четири различите технике машинског учења - *Decision tree, XGBoost, Random Forest* и потпуно повезану неуронску мрежу. Уз то, применили смо *semi-supervised* приступ приликом обучавања модела код два регресиона приступа: класична и *time-series* регресија. Управо примена машинског учења и

статистичких метода за предикцију животног века омогућава боље разумевање овог сложеног феномена и пружа могућност доношења прецизнијих закључака о здравственом стању популације и унапређењу интервенција.

VIII. Литература

[1] Nigri, A., Levantesi, S., & Aburto, J. (2022). Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth. Demographic Research, 47, 199–232.

[2] Ruan, X., Li, Y., Jin, X., Deng, P., Xu, J., Li, N., Li, X., Liu, Y., Hu, Y., Xie, J. and Wu, Y., 2021. Health-adjusted life expectancy (HALE) in Chongqing, China, 2017: An artificial intelligence and big data method estimating the burden of disease at city level. The Lancet Regional Health-Western Pacific, 9, p.100110.

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

[4] World Health Organization, 2020. World health statistics 2020.

[5] 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.

[6] 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.

[7] H. Fan, F. Zhang, R. Wang, X. Huang and Z. Li, "Semi-Supervised Time Series Classification by Temporal Relation Prediction," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 3545-3549, doi: 10.1109/ICASSP39728.2021.9413

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