Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.9
DOI: https://doi.org/10.24867/12BE49Makaric

ПРИМЕНА АЛГОРИТАМА МАШИНСКОГ УЧЕЊА У ПРЕДИКЦИЈИ СТОПЕ САМОУБИСТАВА У ПОЈЕДИНИМ ДРЖАВАМА СВЕТА

THE USE OF MACHINE LEARNING ALGORITHMS IN THE PREDICTION OF SUICIDE RATES IN SOME COUNTRIES OF THE WORLD

Милица Макарић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Самоубиство је сложена појава која вековима привлачи пажњу разних научника. Сваке године, самоубиство је међу 10 водећих узрока смрти у свету међу људима свих узрасних доби. Као озбиљан здравствени проблем захтева нашу пажњу, премда његова контрола и превенција нису нимало једноставни обзиром да на њих могу утицати разни фактори. Из тог разлога, јављају се разне мере које се предузимају у циљу покушаја превенције. Један од тих покушаја спроводи се коришћењем рачунара, односно машинског учења, и разних алгоритама како би се извршила предикција и омогућило спречавање самоубистава у свету. Управо из тога произилази мотивација за овај рад. У раду је представљена предикција стопе самоубистава у различитим државама света. Предикција је извршена помоћу алгоритама машинског учења. Коришћено је више регресионих алгоритама, као и један класификациони алгоритам.*

Кључне речи: *Предикција, Стопа самоубистава, Алгоритми машинског учења*

Abstract – *Suicide is a complex phenomenon that has attracted the attention of various scientists for centuries. Every year, suicide is among the 10 leading causes of death in the world among people of all ages. As a serious health problem, it requires our attention, although its control and prevention are not at all simple, considering that they can be influenced by various factors. For this reason, various measures are being taken to try to prevent it. One of these attempts is carried out using computers, i.e. machine learning, and various algorithms in order to make predictions and enable the prevention of suicides in the world. This is precisely the motivation for this paper. The paper presents the prediction of suicide rates in different countries of the world. Prediction was performed using machine learning algorithms. Several regression algorithms were used, as well as one classification algorithm.*

Keywords: *Prediction, Suicide rates, Machine learning algorithms*

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ванр. проф.

1. УВОД

Од почетка људске цивилизације постојала је потреба да се пронађе начин како се супротставити аутодеструктивном понашању и како повећати осећај задовољства животом. Светска здравствена организација самоубиство дефинише као чин намерног прекида сопственог живота [1]. Према класичним психијатријским ставовима, суицид је чин повезан са поремећајем нагона за самоодржањем [2].

Поставља се питање да ли је могуће направити предикцију степена самоубиства у различитим земљама, како би се у што већој мери утицало на њихово спречавање. Управо тим проблемом се бави овај рад. У њему ће бити приказано једно решење за процену предикције броја самоубиства у одређеним земљама. Подаци на основу кога је обучен модел представљају праћење броја самоубистава од 1990. до 2016. године. Поред броја самоубистава, скуп података садржи и податке о величини популације у држави, старосно доба, пол, просечна примања. У бројним студијама о суициду пронађена је повезаност суицида и сезонских варијација. Изражен је јасан утицај годишњих доба на суициде, са пиком у пролеће и лето у односу на остала годишња доба, што потврђује да стопа суицида током времена кореспондира са сезонским варијацијама [3]. Из тог разлога је скуп података проширен податком о просечном броју сунчаних сати у години.

2. ПРЕГЛЕД ПОСТОЈЕЋИХ РЕШЕЊА

Предикција самоубистава представља предмет многобројних истраживања. Научни радови који се баве овом тематиком разликују се по избору алгоритама за машинско учење, као и по избору фактора на основу којих се врши предикција. У наставку су наведени они радови који су имали највећи утицај на овај рад.

У раду [4] је извршена предикција ризика покушаја самоубистава кроз време коришћењем различитих техника машинског учења. У овом раду су за алгоритме предикције одабрани логистичка регресија (енг. *Logistic regression*), и алгоритам случајне шуме (енг. *Random Forest*). Предикција је вршена на основу медицинских података из клиничких здравствених картона у којима су забележени случајеви повреда за које се зна или сумња да су самонанешене. Такође, у обзир су узети и здравствени картони у којима нису забележени подаци о покушајима самоубиства. Модел је показао добре резултате са следећим вредностима

коришћених евалуционих параметара: АУЦ = 0.84, прецизност = 0.79, одзив = 0.95, и успео је да тачност погађања побољша са 720 дана на 7 дана пре покушаја самоубиства.

Рад [5] се бави предикцијом идеја појединаца о самоубиству на основу социо-демографских, физичких и психолошких обележја. Предикција се врши употребом *Random Forest* алгоритма. Као неки од најбитнијих параметара су издвојени: депресија, ниво стреса у свакодневном животу, пол, старост. Модел предвиђања постигао је добре перформансе (AUC = 0,85) са следећим вредностима валидационих параметара: *accuracy* = 0,821, *sensitivity* = 0,836 и *specificity* = 0,807.

Рад [6] истражује тезу да на суицидно понашање утиче сунчева светлост. За истраживање су коришћени подаци о свим самоубиствима у Аустрији у периоду од 1970-2010. године. Подаци о просечном броју сунчаних сату у току дана добијени су са 86 репрезентативних метеоролошких станица. Као резултат овог истраживања изведен је закључак да постоји позитивна корелација између броја самоубистава и броја сунчаних сати на дан самоубиства, као и до 10 дана пре самоубиства.

3. ПРИКУПЉАЊЕ И ПРИПРЕМА ПОДАТАКА

У овом поглављу је детаљно описан скуп података који је коришћен у пројекту. Наведени су полазни скупови података, као и све трансформације које су над њима примењене како би настао скуп података који је даље коришћен у раду. Такође, представљене су и статистичке анализе које су извршене над поменутим подацима.

3.1. Редукција сувишних обележја

Ово поглавље описује поступак уклањања сувишних обележја. Иницијални скуп података представљао је скуп података преузет са веб странице која садржи велики број различитих скупова података. У питању је сајт *kaggle.com* [7]. У питању је скуп података који поред социо-економских информација, садржи и информације о броју самоубистава за 101 државу. Информације се односе на период од 1985. до 2016. године. Анализом овог скупа података је утврђено да за атрибут *индекс хуманог развоја* велики број редова има недостајуће вредности. Из тог разлога је ова колона изузета из скупа података. Даљом анализом је утврђено да постоји још неколико колона које не доприносе даљем коришћењу скупа података. Једна од тих колона јесте *назив генерације*. Наиме, у скупу података постоји колона *старосна група*, и за сваку старосну групу постоји одговарајући назив генерације. Како су ове две колоне у потпуној корелацији, донета је одлука да се колона *назив генерације* уклони из скупа података. Колона назив државе са годином је такође искључена из даљег разматрања. Вредности које ова колона садржи представљају конкатениране, односно спојене вредности колона *назив државе* и *година*. Из тога се лако закључује да колона назив државе са годином представља редундантан податак, те је то разлог њеног уклањања. Анализом скупа података је донета одлука да се и колона *бруто домаћи производ по глави становника* уклони. Наиме, вредности које ова колона садржи се могу добити дељењем вредности колоне *бруто домаћи производ* са вредностима колоне *број становника државе*.

3.2. Спајање скупова података

На веб страници Економске комисије за Европу (*UNECE*) је пронађен скуп података који садржи информације о просечним месечним примањима држава кроз године [8]. Конкретно, овај скуп података садржи податке о просечној месечној заради за 56 различитих држава у периоду од 1990. до 2017. године. Обзиром да први скуп података садржи информације о 101 држави, а други скуп о 56 држава, приликом спајања ова два скупа података је утврђено да пресек ова два скупа чине 44 различите државе. Поред тога, ова два скупа се односе и на различит опсег година. Због тога је урађен пресек, те се финални скуп података односи на податке између 1990. и 2016. године.

Како је пронађен податак да је највећи број самоубистава забележен у периоду пролеће-лето [3], у разматрање је узет и скуп података о броју сунчаних сати по години у различитим градовима широм света. Адекватни скуп са подацима о сунчаним сатима у току године по државама није пронађен, те су подаци ручно преузети и обрађени 12 са веб странице wikipedia.org [9]. Дати подаци су били груписани по континентима, и приказани за сваки месец у години. Из тог разлога је, након преузимања података, извршено сабирање броја сунчаних сати по месецу за град сваке државе која постоји у циљном скупу података. На тај начин је добијен податак о броју сунчаних сати за сваку државу циљног скупа података.

Овако формирани скуп података је подељен на тренинг и тест подскупове у односу 75% - 25%. Подела је извршена тако да се у тренинг скупу налазе подаци од 1990. до 2008. године, док су у тест скупу подаци везани за период од 2009. до 2016. године. Овако формиран тренинг скуп података је употребљен за обучавање модела како би се извршила предикција стопе самоубиства.

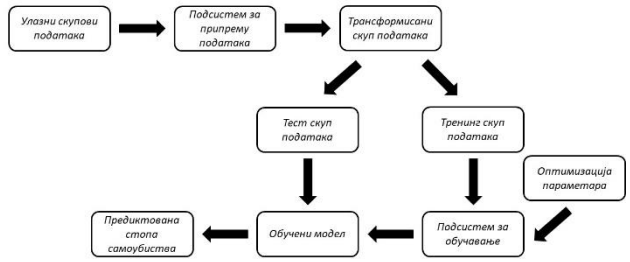
3.3. Трансформације над подацима

Обзиром да су вредности атрибута држава (*country*), пол (*sex*) и година (*age*) у оригиналном скупу података били у текстуалном облику, над овим подацима је било неопходно применити трансформацију у бројеве. Конкретно, назив сваке државе је замењен индексом те државе у скупу података. На тај начин су надаље државе биле представљене бројевима од 0 до 43. Атрибути *sex* и *age* су такође измењени на сличан начин као и атрибут *country*. Вредности атрибута који представља пол су након трансформације садржали само вредности *0* и *1*, уместо вредности *female* и *male*. Атрибут *age* који је садржао 6 текстуалних вредности за 6 различитих старосних група, након промене је презентован вредностима *0-5*.

4. МЕТОДОЛОГИЈА И АЛАТИ

У овом поглављу је представљена методологија која је примењена за предикцију стопе самоубистава над

претходно описаним скупом података. Први део ове целине је задужен за приказ структуре система, заједно са свим подсистемима које обухвата. Обзиром да је у претходном поглављу детаљно описан први подсистем, у овом поглављу ће акценат бити стављен на други подсистем целокупног система – подсистем за обучавање. Подсистем за обучавање се састоји из различитих алгоритама машинског учења. У оквиру овог система је за предикцију стопе самоубистава одабрано неколико различитих регресионих модела, као и један класификациони модел. На слици 1 се налази скица структуре читавог система.



Слика 1. *Шематски приказ структуре система*

4.1. Регресиони модели подсистема за обучавање

Централни део истраживања овог рада фокусиран је на испитивање перформанси различитих алгоритама машинског учења приликом 22 предикције стопе самоубистава. Анализом радова који се баве сличном тематиком као и овај рад, утврђено је који су алгоритми оптимални за овај проблем, и они су и анализирани у наставку.

4.1.1. Линеарна регресија

Линеарна регресија представља један од најједноставнијих и најчешће коришћених модела машинског учења. Односи се на сваки приступ моделовања између једне или више зависних променљивих (означених са *Y*), и једне или више независних променљивих (означених са *X*), на такав начин да модел линеарно зависи од непознатих параметара процењених из података. Обзиром да је у овом раду циљ предикција, линеарна регресија се користи за подешавање предиктивног модела према разматраном скупу података вредности *Y* и *X*. Након развоја оваквог модела, уколико је дата вредност за *X* без припадајуће вредности за *Y*, модел се може употребити за предвиђање вредности *Y*.

4.1.2. Support Vector Regression

Метод потпорних вектора (енг. *Support Vector Machines*) је алгоритам који је првобитно осмишљен за решавање класификационих проблема. Тек касније је овај алгоритам адаптиран за регресионе проблеме. У том случају се користи назив регресија потпорних вектора (енг. *Support Vector Regression*). Алгоритам SVM је заснован на једноставној идеји, а то је да се дефинише хиперраван која раздваја податке који припадају различитим класама. Кључна разлика између метода потпорних вектора за класификацију и за регресију је та што се у случају регресије не траже тачна предвиђања, као што је у линеарно раздвојивом случају код класификације био захтев. У пројекту који је описан у овом раду је за алгоритам *SVR* извршена оптимизација два параметра: параметар C (мера којом се пенализује грешка) и избор кернел функције. У овом примеру се као најпогоднија показала *RBF* кернел функција. Емпиријски је утврђено да оптимална вредност параметра C за претходно описани скуп података износи 1000.

4.1.3. Gradient Boosted Tree

Алгоритам *Gradient Boosted Tree* се у српском језику може пронаћи под називом *алгоритам појачавања градијената*. Главна идеја код *boosting* алгоритама јесте идеја да се слаб предиктор може модификовати да постане бољи. Слаб предиктор или слаб ученик се дефинише као онај чији је учинак бар мало бољи од случајне шансе. *Појачавање* је ансамбл техника у којој се предиктори не праве независно, већ узастопно. Ова техника користи логику у којој наредни предиктори уче на грешкама претходних предиктора. Будући да нови предиктори уче на грешкама које су починили претходни предиктори, потребно је мање итерација да би се приближили стварним предвиђањима. Као слаб предиктор се користе стабла одлучивања (енг. *decision trees*). У примеру за предикцију стопе самоубиства који је обрађиван у овом раду је за алгоритам *Gradient Boosted Tree* извршена оптимизација неколико параметара: *n_estimators* (број стабала одлучивања) = 100, *lerning_rate* (стопа учења) = 0.4 и *max_depth* (максимална дубина стабла одлучивања) = 4.

4.1.4. Extreme Gradient Boosting

Extreme Gradient Boosting или скраћено *XGBoost* алгоритам је алгоритам који у последње доба доминира примењеним машинским учењем. Овај алгоритам представља имплементацију стабала одлучивања појачаних градијената дизајнираних за рачунарску брзину и перформансе модела. *Extreme Gradient Boosting* је једна од имплементација *Gradient Boosting* концепта, али оно што *XGBoost* чини јединственим јесте то што користи уређенију формализацију модела за контролу прекомерног уклапања, што му даје боље перформансе. Модел алгоритма *XGBoost* који је коришћен у овом раду је оптимизован. Конкретно оптимизовани су параметри: *n_estimators*=200, *learning_rate*=0.16 и *max_depth*=4.

4.1.5. Random Forest

Алгоритам *Random Forest* се у нашем језику преводи као *алгоритам случајних шума*. То је алгоритам надгледаног учења који се користи и за регресију и за класификацију. Случајна шума је метод учења који делује конструисањем вишеструких стабала одлучивања на узорцима података. Овај метод креира велики број стабала одлучивања у време тренирања, у сврху елиминисања шума и аутлајера (енг. *outlier*). Од креираног скупа стабала одлучивања, метод случајне шуме добија предвиђање од сваког стабла појединачно, и на крају бира најбоље решење гласањем. За потребе овог истраживања, алгоритам Random Forest је оптимизован. Оптимизација је урађена за следеће параметре: *n_estimators*=20, *max_depth*=50 и *min_samples_split*=4.

4.2. Класификациони модел подсистема за обучавање

У сврху класификационог модела за предикцију стопе самоубистава којом се бави овај рад, одабран је алгоритам *Random Forest*. Како је његов метод рада описан у претходном поглављу, ово поглавље ће бити усмерено на приказ оптимизационих параметара који су употребљени у класификационом моделу. Оптимизовани су исти параметри као и у случају регресије, и њихове оптимизоване вредности износе: *n_estimators*=20, *max_depth*=15, *min_samples_split*=6.

5. ЕКСПЕРИМЕНТАЛНА ЕВАЛУАЦИЈА

Као мера евалуације перформанси регресионих модела, коришћена је R^2 мера, као и корен средње вредности квадрата грешке (енг. *Root Mean Square Error – RMSE*). У табели 1 су приказане вредности ових параметара које су добијене евалуацијом регресионих модела.

Табела 1. *Резултати регресије на тест скупу*

	R^2	*RMSE*
Linear Regression	0.65	493.4
Support Vector Regression	0.26	704.37
Gradient Boosted Tree	0.97	135.03
XGBoost	0.98	128.46
Random Forest	0.95	179.8

На основу података приказаних у табели 1, закључује се да је за проблем описан у овом раду најбоље резултате постигао алгоритам *XGBoost*. Коефицијент детерминације за овај алгоритам износи 0.98, што је врло близу максималној вредности. Такође, овај алгоритам је постигао и најмању вредност за *RMSE*. Веома добре резултате се постигли и алгоритми *Gradient Boosted Tree* и *Random Forest*, те се и они могу сматрати погодним алгоритмима за овакав проблем.

За проблем класификације је као мера евалуације коришћена F_1 мера, као и параметри прецизност (енг. *precision*) и одзив (енг. *recall*). У табели 2 су приказани резултати евалуације класификационог проблема.

Табела 2. *Резултати класификације на тест скупу*

	F_1	*precision*	*recall*
Random Forest	0.748	0.752	0.75

Као што се из претходне табеле може уочити, вредности сва три разматрана параметра имају приближно исту вредност која износи *0.75*. Обзиром да максимална вредност за ове параметре износи *1*, може се закључити да су резултати добијени класификацијом у великој мери прихватљиви.

6. ЗАКЉУЧАК

У раду је приказан систем за предикцију стопе самоубиства појединих држава света на основу аллгоритама машинског учења. Приликом предикције је коришћено више различитих регресионих модела, као и један класификациони модел. Приликом предикције су, осим основних података о државама, коришћени и одређени специфични фактори попут просечних месечних примања грађана и броја сунчаних сати у години. Решење овог проблема довело би до знања који то фактори доводе индивидуу до ситуације да почини самоубиство, што би могло помоћи у превенцији таквих ситуација.

Скуп података који је коришћен у овом раду садржи податке о 44 различите државе у периоду од 1990. до 2016. године. Овај скуп је подељен на тренинг и тест подскупове у односу 75% - 25%. Након тога је путем регресије и класификације обучено више модела на тренинг скупу података. Најбољи резултати су постигнути приликом употребе регресионог модела *XGBoost*. У том случају је постигнута R^2 вредност од 0.98. Даљи правци развоја овог решења обухватају добављање података о другим државама, добављање података о годинама пре 1990. и након 2016. године, као и коришћење других алгоритама машинског учења за предикцију.

7. ЛИТЕРАТУРА

[1] World Health Organization, „Preventing suicide: A global imperative“, Geneva, 2014.

[2] Д. Марчинко, „Модели разумјевања суицидалнога понашања“, Медицинска наклада Загреб, 2011.

[3] Ч. Милић, „Сезонске варијације – фактор ризика за настанак суицида“, Медицински преглед Нови Сад, 2010.

[4] C. G. Walsh, J. D. Ribeiro, J. C. Franklin, „Predicting Risk of Suicide Attempts Over Time Through Machine Learning“, Florida State University, Florida, 2017.

[5] S. Ryu, H. Lee, D. K. Lee, K. Park, „Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population“, Republic of Korea, 2018.

[6] B. Vyssoki, N. D. Kapusta, N. Praschak-Riede, G. Dorffner, M. Willeit, „Direct Effect of Sunshine on Suicide“, Austria, 2014.

[7] https://www.kaggle.com/russellyates88/suicide-rates-overview1985-to-2016 (приступљено у октобру 2020.)

[8] „Gross Average Monthly Wages by Country and Year”, [Online] Available: https://w3.unece.org (приступљено у октобру 2020.)

[9] https://en.wikipedia.org/wiki/List_of_cities_by_-sunshine_duration (приступљено у октобру 2020.)

Кратка биографија:

Милица Макарић рођена је у Новом Саду 1996. године. Факултет техничких наука у Новом Саду, смер Рачунарство и аутоматика, уписала је 2015. године. Основне академске студије је завршила 2019. године, након чега је уписала мастер академске студије на Факултету техничких наука, смер Рачунарство и аутоматика.
контакт: makaric.milica@gmail.com