Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.85
DOI: https://doi.org/10.24867/20BE08Batinic

АНАЛИЗА ФАКТОРА И ПРЕДИКЦИЈА СТОПЕ САМОУБИСТАВА ПО ДРЖАВАМА БАЗИРАНО НА МОДЕЛИМА ДУБОКОГ УЧЕЊА

FACTOR ANALYSIS AND PREDICTION OF SUICIDE RATES BY COUNTRIES BASED ON DEEP LEARNING MODELS

Николина Батинић, *Факултет техничких наука, Нови Сад*

Област – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *Самоубиство је један од водећих узрока смрти у свијету, што је изузетно забрињавајуће. У овом раду се примјеном техника вјештачке интелигенције врши анализа фактора који утичу на самоубиства, као и предикција стопе самоубиства за различите групације људи. Са тим циљем прикупљено је више скупова података који садрже различите факторе који могу утицати на стопу самоубиства у држави, како би се испитао њихов значај. Примери анализираних фактора су држава, пол, старосна група, година и коефицијент незапослености. Над претпроцесираним подацима испробана су два приступа: (1) класичан приступ (који обухвата више различитих регресионих модела) и (2) и приступ базиран на временским серијама (где се користи LSTM мрежа).*

Кључне ријечи: *временске серије, рекурентне неуронске мреже, LSTM, регресија.*

Abstract – *Suicide is one of the leading causes of death worldwide, which is highly worrying. This paper applies artificial intelligence to analyze the factors that might influence suicide and predict the suicide rate for different groups. Several data sets containing different factors that might influence the countries' suicide rates were collected to analyze their importance. These factors include country, gender, age group, age, and unemployment rate. Multiple prediction approaches were applied to the preprocessed data: (1) a traditional approach (multiple regression models) and (2) a time series approach (using the LSTM model).*

Keywords: *time series, RNN, LSTM, regression*

1. УВОД

Према подацима Свјетске здравствене организације (енг. *World Health Organization* – WHO) у свијету се сваких 40 секунди убије једна особа. У посљедњих пола вијека стопа самоубиства на свјетском нивоу је повећана за 60% [1].

У посљедњих неколико година, медицински стручњаци и истраживачи су се све више фокусирали на препознавање и спрјечавање самоубилачких образаца и понашања. Поред бројних напора стручњака и програма за превенцију самоубистава, и даље не се види значајно побољшање [2]. Међутим, са напретком технологије и вјештачке интелигенције, проблему се може приступити на другачији начин.

У овом раду се примјеном техника вјештачке интелигенције врши анализа фактора који утичу на самоубиства, као и предикција стопе самоубиства за различите групације људи. У овом циљу, у раду се врши:

- Анализа и обрада података – врши се претпроцесирање и издвајају фактори који ће бити укључени у тренирање модела;
- Предикција стопе самоубиства – примјењују се два различита приступа: класичан приступ (регресиони модели) и приступ базиран на временским серијама (енг. *time series*), што су у овом случају панел (лонгитудинални) подаци.

Добијеним резултатима се може допринијети превенцији самоубиства, што представља и мотивацију за израду овог рада.

У класичном, *cross-sectional* приступу, за предикцију је испробано више различитих регресионих модела, попут: регресије методом случајне шуме (енг. *Random Forest*), *Gradient boosting Regressor*-а и *Ridge* регресије. Код методологије базиране на временским серијама кориштена је рекурентна неуронска мрежа (енг. *Recurrent Neural Network* - RNN), LSTM (*Long Short-Term Memory*) архитектуре.

У наредном поглављу дат је преглед претходних рјешења који су се бавили истим или сличним проблемом као и овај рад. Детаљи рјешења и имплементације система су описани у поглављу 3, а поглавље 4 садржи анализу добијених резултата. Посљедње, 5. поглавље, представља закључак рада.

2. ПРЕТХОДНА РЈЕШЕЊА

Проблем проналажења фактора који имају највећи утицај на повећану стопу самоубиства је у вјештачкој интелигенцији присутан већ неко вријеме. У посљедњим годинама се све више користе методе дубоког учења за рјешавање овог проблема.

У раду [3] су анализирани узроци који утичу на стопу самоубиства у Индији. Скуп података који је кориштен чине подаци који садрже информације о самоубиствима у Индији, објављен од стране

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је била др Јелена Сливка, ванр. проф.

Националног бироа за евиденцију криминала у Индији у периоду од 2001. до 2012. године. Неки од фактора који представљају улазне варијабле су: пол, држава у Индији, година, старосна група, начин самоубиства, брачни статус и ниво образовања. Лабела је представљала претпостављени узрок (болест, породични проблеми, итд.).

За предикцију су кориштени вјештачка неуронска мрежа (енг. *Artificial Neural Network* - ANN) и метод потпорних вектора (енг. *Support Vector Machine* - SVM). ANN је дао тачност од 77.5%, а SVM 81.5%. Оно што је значајно за овај рад је да се у раду [3] показало да многи од фактора које и овај рад испитује, имају значајан утицај на ризик од самоубиства, с тим што се предикција вршила само за подручје Индије и што подаци нису посматрани као временске серије.

Рад [4], у коме се користи скуп података који је један од скупова података који чине финални скуп у овом раду, пореди ефикасност различитих модела машинског учења за предвиђање стопе самоубиства у зависности од државе. Модели су обучавани на дијелу података Свјетске здравствене организације, који садржи преглед стопе самоубиства од 1985. до 2016. године. Независне промјењиве кориштене за обучавање су: држава, година, бруто домаћи производ за ту годину, пол, старост, генерација. Испробан је велики број различитих модела, а најслабије резултате су дали *Ridge*, LASSO и *Elastic net* регресија. Значајно бољи резултати су добијени уз помоћ *Decision Tree Regressor*-а и регресије методом случајне шуме. У овом раду су испробане наведене методе из рада [4]. Главни недостатак рада [4] је то што није вршена експлоративна анализа података и што није испробан приступ базиран на временским серијама.

Када су у питању предикције временских серија, у раду [5] се пореди ефикасност обичне рекурентне неуронске мреже и LSTM-а приликом предикције вриједности акција са подацима базираним на временским серијама. Кориштена су два независна скупа: дневни подаци о тржишту са Шангајског композит индекса и са Дау Џонс индекса у периоду од 2007. до 2017. године. Почетна цијена, цијена затварања и најнижа дневна цијена представљају независне промјењиве, док је цијена затварања следећег дана зависна промјењива. За обучавање модела кориштени су подаци из првих осам година, док су за тестирање кориштени подаци из посљедње двије године из скупа података. LSTM је показао боље резултате у односу на RNN. Из тог разлога, LSTM је испробан и у овом раду, иако се ради са панел подацима.

3. МЕТОД

У наредним поглављима су описани скупови података кориштени у раду, начин на који је извршена експлоративна анализа података и архитектура испробаних модела за предикцију.

3.1. Скупови података

Финални скуп података, над којим су тренирани и тестирани сви кориштени модели у овом раду је креиран анализом, интеграцијом и трансфомрацијом више различитих јавно доступних скупова података.

Први скуп података [6] чини укупно 27.820 инстанци. Циљну лабелу представља број самоубистава на 100.000 становника. Остала обиљежја су: држава, година, пол, старосна група, број самоубистава, популација, индекс хуманог развоја (енг. *Human Development Index* – HDI), бруто домаћи производ (енг. *Gross Domestic Product* – GDP) по становнику и укупно и демографска генерација којој припада популација. Скуп садржи информације од 1987. до 2016. године уз велики број недостајућих вриједности за неке од атрибута, попут HDI-а.

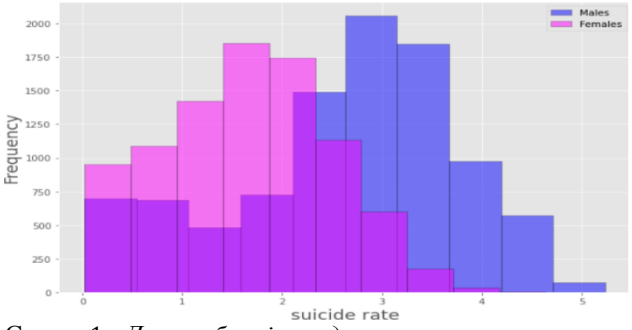
Други скуп података, који је кориштен у раду је преузет са сајта Свјетске банке (енг. *World bank*) [7]. Садржи информације о основним економским параметрима већина држава свијета у периоду од 1985. до 2016. године. Неке од информација које представљају обиљежја су: раст GDP-а по становнику, коефицијент незапослености по половима и ЏИНИ коефицијент. За преко 6.000 података недостају информације о ЏИНИ коефицијенту, због чега је кориштен и трећи скуп података [8] који садржи информације о остваренoм ЏИНИ индексу за све државе по годинама (од 1800. до 2016. године).

У раду је испробан и четврти скуп података [9], који садржи информације о стопи развода по годинама и државама. Укупно има 2.819 података које чине информације од 1991. до 2018. године. За велики број држава и година нема података.

3.2. Експлоративна анализа и обрада података

Над претходно описаним скуповима података извршена је експлоративна анализа и обрада. Овај процес је врло често круцијалан у поступку предикције, јер је то једини начин да сирови подаци постану употребљиви. Први корак представља анализа и обрада сваког од скупова података понаособ.

У првом скупу података, који између осталог садржи и циљну лабелу, уклоњене су колоне које садрже велики број недостајућих вриједности, као и колоне које су у савршеној корелацији са неким од осталих обиљежја (нпр., укупан GDP и GDP по становнику). На слици 1 је приказана дистрибуција података на основу стопе самоубистава у зависности од пола, након што је примијењена логаритамска трансформација циљне лабеле. Оно што се најприје може закључити је већа стопа самоубистава код мушкараца, него код жена.



Слика 1. *Дистрибуција података на основу стопе самоубистава у зависности од пола.*

Из другог скупа података су издвојене колоне које представљају незапосленост по половима, те су уклоњени редови који садрже недостајуће вриједности на основу овог обиљежја. Након тога је овај скуп података интегрисан са претходно објашњеним скупом на основу државе и године на коју се односе подаци. Трећи скуп података, који садржи информације о ЏИНИ индексу држава по година, је интегрисан са претходна два.

На финални скуп, који чине подаци из претходно наведених скупова података, додат је и четврти скуп који садржи информације о стопи развода. Након ове интеграције, дошло је до значајног губитка података, због чега су за предикцију испробана два финална скупа података (са и без стопе развода). Из финалних скупова избачене су државе које су садржале велики број података са стопом самоубиства која износи 0, због ирелевантности информација. На крају, финални скуп података, без стопе развода, је имао 17.429 узорака, док је други скуп, који садржи и стопу развода, имао око 10.000 података.

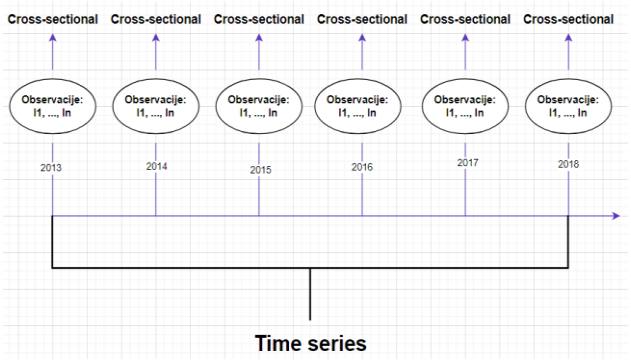
Након интеграције и креирања финалног скупа података, поново је извршена анализа, на основу које се могло закључити да у женској популацији, на глобалном нивоу, постоји благи тренд опадања стопе самоубистава, док у мушкој популацији ова вриједност варира кроз вријеме. Осим тога, закључује се и да највећу стопу самоубистава има Највећа генерација (енг. *Greatest generation*), што је и очекивано, с обзиром да ову генерацију чине дјеца Првог свјетског рата и борци Другог свјетског рата, са бројним траумама.

3.3. Архитектура рјешења

За предикцију стопе самоубиства кориштена су два различита приступа: класичан (*cross-sectional*) приступ и приступ заснован на панел подацима.

Класичан приступ подразумијева да се подаци посматрају као опсервације различитих ентитета у једном временском тренутку. Вријеме, што је у овом случају година, представља категоричко обиљежје, као и свако друго, што значи да се не прати тренд кретања вриједности циљне лабеле кроз вријеме. Над њим, као и осталим категоричким обиљежјима, извршена је конверзија у нумеричку вриједност, примјеном *label encoding*-а. За предикцију је кориштено више различитих модела машинског учења за регресију. Испробани су следећи модели: *Ridge*, LASSO и *Elastic net* регресија, *Decision Tree Regressor*, регресија методом случајне шуме и *Gradient Boosting Regressor*.

Финални скуп података, којим се бави овај рад, се може посматрати и као скуп лонгитудиналних или панел података (комбинација временских серија и *cross-sectional* података). Разлог је што у једној години постоје подаци за више држава, док истовремено за једну државу постоје и подаци кроз вријеме, са временским интервалом од једне године (слика 2). Из разлога што постоји више ентитета, што су у овом случају државе, подаци се не могу посматрати као класичне временске серије.



Слика 2. *Илустрација панел дизајна*.

Код регресије панел података је битно да постоји једна колона која јединствено идентификује ентитет. С обзиром на то да у финалном скупу података овог рада за сваку државу постоје опсервације по половима и старосној групи којој та група људи припада, идентификатор представља нова колона, коју чини комбинација следеће три колоне: држава, пол и старосна група. Идентификатор ентитета и колона која представља временску инстанцу, што је у овом случају година, заједно чине *multi-index*.

Прије примјене LSTM-а, извршено је додатно претпроцесирање података како би се подаци припремили за примјену модела. Први корак је „*Lag timestamp*“. Он представља трансформацију *time series* података у *dataframe*, гдје ће сваки податак, осим информација о тренутној опсервацији, садржати и вриједности обиљежја из претходне опсервације, груписан по индексу. Након овога, подаци су додатно преобликовани у 3D формат, да би представљали улаз у LSTM мрежу. Архитектура мреже која је кориштена у овом раду је *Vanilla* LSTM архитектура која се састоји од три слоја: улазни, скривени и један излазни *Dense* слој. Функција губитка (енг. *loss function*) која је кориштена при компајлирању модела је MAE (*Mean Absolute Error*). Модел је трениран у 10 епоха, а *batch size* износи 16.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Када је у питању класичан приступ подацима за регресију, скуп података је подијељен на тренинг и тест у размјери 80:20. За оптимизацију параметара модела кориштена је унакрсна валидација (енг. *cross validation*). Мјере евалуације које су кориштене за одређивање перформанси модела су: R2 (*R-Squared*) и RMSE (*Root Mean Square Error*).

Модел	R2	RMSE
Ridge регресија	0.29	0.94
LASSO регресија	0.29	0.94
Elastic net регресија	0.29	0.94
Decision Tree Regressor	0.89	0.36
Random Forest Regressor	0.64	0.67
Gradient Boosting Regressor	**0.94**	**0.26**
***Gradient Boosting Regressor* (2. скуп)**	**0.96**	**0.23**

Табела 1. *Перформансе cross-sectional модела*.

У табели 1 су дати резултати које су остварили испробани модели над тест скупом података у *cross-sectional* приступу. Као што се може видјети, најбоље резултате је дао *Gradient Boosting Regressor*. Исти модел је испробан и над другим скупом података, који је садржао и информације о стопи развода по државама и годинама. Остварени резултати су за нијансу бољи. Због тога се може закључити да овај фактор има утицај на повећање стопе самоубистава, иако је број података доста мањи.

У приступу базираном на временским серијама експеримент је постављен на више различитих начина. Први начин подразумијева предикцију стопе самоубиства за посљедњу годину за коју постоје подаци. У овом случају, то ће бити 2015. година, јер је 2016. изузета из скупа података, због великог броја недостајућих вриједности. Вриједност додатног параметра, који представља број претходних година чији подаци чине тренинг скуп, као и вриједност осталих хипер-параметара је оптимизована на основу валидационог скупа, кога чине подаци из 2014. године. Други приступ подразумијева да се резултати предвиђају за посљедње двије године из скупа података, док валидациони скуп чине двије године прије. Посљедњи начин на који је постављен експеримент подразумијева подјелу скупа на тренинг и тест у размјери 80:20, временски гледано. У оваквом приступу су због хардверских ограничења кориштени параметри LSTM модела добијени из другог приступа. У табели 2 су дати добијени резултати за сваки од постављених експеримената.

Модел	R2	RMSE
LSTM (2015. година)	0.16	0.76
LSTM (посљедње 2 године)	0.32	0.45
LSTM (посљедњих 20% година)	**0.71**	**0.30**

Табела 2. *Перформансе time series модела*.

У оваквом приступу најбољи резултат је остварио модел који на основу првих 80% података, предвиђа резултате за посљедњих 20%. Битно је напоменути да није у потпуности релевантно поредити резултате ових модела, јер је за сваки од њих тест скуп другачији.

5. ЗАКЉУЧАК

Овај рад се бави анализом фактора који утичу на стопу самоубистава у свијету, праћену кроз вријеме по државама, као и самом предикцијом стопе самоубистава. Мотивацију је представљало то што би се добијеним резултатима могло допринијети превенцији великог броја самоубистава, јер је то један од водећих узрока смрти у свијету.

За предикцију су испробана два различита приступа: *cross-sectional* приступ, који не узима у обзир временску зависност података и приступ базиран на панел подацима. У другом приступу примјењен је LSTM модел рекурентних неуронских мрежа. Експеримент је постављен на више различитих начина и упоређење су перформансе за сваки од њих. Финални скуп података, кориштен за тренирање и тестирање модела, чини комбинација више скупова података, прикупљених из различитих извора. Код класичног приступа, најбољи резултат је остварио *Gradient Boosting Regressor*. У приступу базираном на временским серијама гдје је примјењен LSTM модел, најбољи резултати је остварен у експерименту у коме се на основу првих 80% података предвиђају резултати за посљедњих 20%.

Предмет даљих истраживања би могла бити анализа и укључивање додатних фактора, попут вјере, образовања, учесталост коришћења друштвених мрежа итд. Осим тога, потенцијално побољшање се може постићи примјеном некох комплекснијег RNN модела, попут GRU (*Gated Recurrent Unit*) или примјеном неког од трансформер модела.

6. ЛИТЕРАТУРА

[1] Ferretti, Fabio & Coluccia, Anna. (2009). Socio-economic factors and suicide rates in European Union countries. Legal medicine (Tokyo, Japan). 11 Suppl 1. S92-4. 10.1016/j.legalmed.2009.01.014.

[2] J. D. Ribeiro, J. C. Franklin, K. R. Fox, K. H. Bentley, E. M. Kleiman, B. P. Chang, and M. K. Nock, "Self-Injurious Thoughts and Behaviors as Risk Factors for Future Suicide Ideation, Attempts, and Death: a Meta-Analysis of Longitudinal Studies," Psychological Medicine, vol. 46, no. 2, pp. 225–236, 2016. DOI: 10.1017/S0033291715001804

[3] Imran Amin, Sobia Syed, Prediction of Suicide Causes in India using Machine Learning, Journal of Independent Studies and Research (JISR), Volume 15, Issue No 2, 2017.

[4] Jhansi lakshmi Durga Nunna, Akila Rani M., B. V. Ram Kumar, Design of Machine Learning based Suicide Rate Prediction System, International Journal of Scientific Research and Review, Volume 8, Issue 4, 2019

[5] Qiang Jiang, Chenglin Tang, Stock price forecast based on LSTM neural network, International Conference on Management Science and Engineering Management, Springer, pp.393-408, 2018.

[6] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

[7] https://databank.worldbank.org/source/world-development-indicators

[8] https://www.kaggle.com/psterk/income-inequality?select=gini.csv

[9] https://ourworldindata.org/marriages-and-divorces

Кратка биографија:

Николина Батинић рођена је 1997. године у Милићима, БиХ. Основне академске студије завршила је 2020. године на Факултету техничких наука, на ком брани и мастер рад 2022. године из области Електротехнике и рачунарства – Софтверско инжењерство и информационе технологије.
контакт: nina.batinic@yahoo.com