Предлог пројекта

Системи за истраживање и анализу података

Овај документ садржи кратак опис теме пројекта и дефиниција, мотивације за одабрану тему и опис проблема. Након што је дат увид у мотивацију проблема приказан је преглед других литература на сличну тему, скуп података и метод евалуације.

Дефиниција пројекта

Задатак рада представља предикцију стопе самоубиства у изабраним државама света на основу досадашњег тренда и социоекономских података. Подаци који указују на досадашњи тренд односе се на период од 1990. до 2016. године. На основу добијеног модела, биће вршено предвиђање стопе самоубиства за 2017. и 2019. годину за сваку од држава из скупа података.

Мотивација

Самоубиство је сложена појава која вековима привлачи пажњу разних научника. Сваке године, самоубиство је међу 20 водећих узрока смрти у свету међу људима свих узрасних доби. Као озбиљан здравствени проблем захтева нашу пажњу, премда његова контрола и превенција нису нимало једноставна како на њих могу утицати разни фактори. Из тог разлога, стално постоје 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.

Преглед других литература

[1] Colin G. Walsh, Jessica D. Ribeiro, Joseph C. Franklin (2017) *Predicting Risk of Suicide Attempts Over Time Through Machine Learning*
https://journals.sagepub.com/doi/abs/10.1177/2167702617691560

Задатак рада: Предикција ризика покушаја самоубиства кроз време коришћењем машинског учења. Примењени су новији и традиционалнији начини коришћења машинског учења како би се показао напредак технологија али и бољи односно прецизнији резултат коришћењем новијих технологија.
Методологија: Коришћена је логистичка регресија и random forest метод.

Подаци: Одређени део података су медицински подаци из репозиторијума клиничких здравсених картона из медицинског центра са Вандербилт универзитета где су забележени случаји повреда за које се зна или сумња да су самонанешене. Други део података је везан је за случајан скуп здравствених картона за које нису забележени подаци покушаја самоубиства. Од параметара рад је узео у обзир демографске податке попут година, пола, расе/нациналности, ранијих дијагноза, социо-економског статуса где су узети образовање, имовина и подаци о запослености.
Евалуација решења: За евалуацију модела коришћена је AUC (area under the receiver operating characteristic curve) као и *precision* и *recall* метрике. Поред тога, за калибрацију унутар самог рада коришћен је и Бријеров резултат (енгл. *Brier score*).

Модел је показао добре резултате са прецизношћу од 0.79 и успео је да прецизност погађања сведе са 720 дана на 7 дана пре покушаја самоубиства. Из тог разлога овај рад показује успешност *random forest* алгоритма као и ефикасан начин евалуације решења приликом имплементације нашег рада.

[2] Seunghyong Ryu, Hyeongrae Lee, Dong-Kyun Lee, and Kyeongwoo Park (2018) *Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population*
https://www.researchgate.net/publication/328225907_Use_of_a_Machine_Learning_Algorithm_to_Predict_Individuals_with_Suicide_Ideation_in_the_General_Population

Задатак рада: Развој модела који предвиђа идеје појединаца о самоубиству унутар опште популације користећи алгоритам машинског учења.
Методологија: Коришћен је Random Forest алгоритам.
Подаци: Од првобитних 47 социо-демографских, физичких и психолошких обележја који утичу на идеју о самоубиству, рекурзивном методом је одабрано 39 карактеристика за које је примећено да модел постиже највећу тачност употребом унакрсне валидације. Како би се смањила димензионалност модела, скуп је смањен на 15 обележја за које се показало да дају незнатно ниже резултате процене од претходно одабраних 39 обележја. Најзначајније карактеристике од коришћених су: депресивно расположење током две недеље, ниво стреса у свакодневном животу, анксиозност/депресија, пол, образовање, субјективно здравствено стање, старост.
Евалуација решења: Скуп података од укупно 11.628 појединаца (5.814 појединаца са идејом о самоубиставу и исто толико појединаца који нису имали идеје о самоубиству) подељен је на тренинг (10.466 појединаца) и тест скуп (1.162 појединаца) уз очување односа 1: 1 између две класе. Евалуација модела је урађена коришћењем AUC (*area under the receiver operating characteristic (ROC) curve*). Такође, из матрице конфузије израчунати су параметри тачност (*accuracy*), осетљивост (*sensitivity*), специфичност (*specificity*), позитивну предиктивну вредност и негативну предиктивну вредност.

Коришћена је 10-кратна унакрсна валидација да би се избегло прекомерно уклапање и повећала уопштеност модела.
Модел предвиђања постигао је добре перформансе (AUC = 0,85) у тестном сету и предвидео идеје појединаца о самоубиству међу укупним узорцима са следећим вредностима валидационих параметара: *accuracy* = 0,821, *sensitivity* = 0,836 и *specificity* = 0,807. Из претходно изложеног, уочава се погодност коришћења Random Forest алгоритма за решавање проблема овог пројекта. Овај рад би било добро допунити са другим алгоритмима машинског учења, како би се упоредиле перформансе модела предвиђања, као што метод потпорних вектора (SVM) и вештачке неуронске мреже.

[3] Sung Man Bae, Seung A Lee, Seung-Hwan Lee (2015) *Prediction by data mining, of suicide attempts in Korean adolescents: a national study*
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4577255/

Задатак рада: Развој модела за предвиђање покушаја самоубистава јужнокорејских адолесцената.
Методологија: Data mining на основу стабла одлучивања, употребом Answer Tree 3.0 софтвера. Као критеријум поделе у оквиру стабла одлучивања кориштена је статистичка техника CHAID (Chi-square Automatic Interaction Detector) с циљем да се на основу међусобних веза великог броја променљивих оптимално формира структура стабла.
Подаци: Студија се базира на истраживању Корејског националног иститута за политику младих, спроведеном 2011. године. Скуп података композицију социодемографских и психолошких података испитаника (степен депресије, стрес, делинквенција, самопуздање, оптимизам) и податка да ли је испитаник у претходних годину дана покушао да изврши самоубиство.
Евалуација решења: Подаци су подељени у тренинг и тест скуп у односу 70/30. Валидација на основу тестног скупа је показала прецизност од 0.9 и стандардну грешку у вредности од 0.00671728, те је закључено да је предикциони модел могуће генерализовати.

Data mining овог истражвања ослања се искључиво на стабло одлучивања. Стабло одлучивања са собом повлачи опасност од потенцијалног оверфитовања модела у случајевима када је присутна велика количина података. Међутим, због чињенице да је у овом истраживању укупан број испитаника био свега 2754, та могућност је мала. Исто тако, велики број променљивих може да утиче на дубину стабла, што отежава класификацију. Овај потенцијални проблем могао би се превазићи употребом Random Forest алгоритма.

Скуп података

Један скуп података преузет је са сајта Kaggle (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016) који садржи податке од 1985. до 2016. године о називу државе, години, бруто домаћем производу држава за сваку годину, бруто домаћем производу по глави становника, полу, рангу година, броју самоубистава, броју становника, броју самоубистава на 100 хиљада становника, индексу хуманог развоја и називу генерације у односу на ранг година. Постоје обележја за која сматрамо да неће имати допринос раду те су та обележја избачена. Конкретно, то су следећи подаци: бруто домаћи производ по глави становника, број самоубистава и назив генерације у односу на ранг година. У овом скупу података постоје подаци за 101 државу, при чему је укупан број редова 27821. Подаци су дати у csv датотеци.

Други скуп података преузет је са UNECE сајта (https://w3.unece.org/PXWeb2015/pxweb/en/STAT/STAT__20-ME__3-MELF/60_en_MECCWagesY_r.px/) и садржи податке о просечним месечним приходима по години и држави од 1990. до 2017. године. Скуп података приказује подаке о 52 државе света. У овом скупу података постоји укупно 1456 редова. Подаци су преузети као csv датотека.

Како би се претходна два скупа података повезала начињене су одређене измене. Посматрају се подаци за 52 државе које постоје у другом наведеном скупу података и просечна примања становника тих држава у периоду од 1990. до 2016. године. Подаци ова два скупа података ће се спајати на основу два обележја: називу државе и години за коју су везани остали подаци.

Методологија

Користиће се Random Forest алгоритам који разматра два параметра: број атрибута које стабло треба да размотри и број стабала који ће бити изграђен. Овај алгоритам је отпоран на оверфитовање, те је претпоставка да ће имати висок проценат тачности. Такође ће се користити и Support Vector Machine (SVM) као и неуронска мрежа. Добијене резултате бисмо упоредили како би пронашли оптималан случај за предикцију. На почетку ће бити извршено претпроцесирање података, како би предикција била што успешнија. Наиме, као најважнија обележја ће рекурзивном методом бити издвојена она за које се покаже да модел има највећу тачност употребом унакрсне валидације. С обзиром да нема много обележја, редукција димензионалности се неће радити.

Метод евалуације

Скуп података ће се поделити на тренинг и тест скуп. Тренинг скуп ће обухватати податке од 1990. до 2008. године, док ће тест скуп обухватати податке од 2009. до 2016. године. За евалуацију резултата користиће се параметар *accuracy,* матрица конфузије заједно са параметрима *precision* и *recall*.

Тим

Тим чине: Вукашин Јовић (E2 59/2019), Милица Макарић (E2 55/2019) и Милан Лазић (E2 73/2019)