Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.9
DOI: https://doi.org/10.24867/25BE18Jeremic

ПРЕДИКЦИЈА СРЧАНОГ УДАРА НА ОСНОВУ СТАЊА ПАЦИЈЕНТА ПРИМЕНОМ АЛГОРИТАМА МАШИНСКОГ УЧЕЊА

HEART ATTACK PREDICTION BASED ON THE PATIENT'S CONDITION USING MACHINE LEARNING ALGORITHMS

Андријана Јеремић, Јелена Сливка, *Факултет техничких наука, Нови Сад*

Област – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *Срчани удар или акутни инфаркт миокарда је изумирање дела срчаног мишића због наглог престанка циркулације кроз неку од артерија које исхрањују срце. Један је од најчешћих узрока смрти у развијеним земљама и земљама у развоју. Чешће погађа мушкарце него жене, а после уласка у климакс, ризик од настанка болести се изједначава. Присутнији је код људи старијих од 40 година. Фактори кардиоваскуларних ризика су различити и на неке од њих се може утицати, а на неке не. Предикција срчаног обољења код особе може бити врло непредвидива и дуготрајна, а често и неуспешна. Зато је данас све заступљенија употреба вештачке интелигенције за решавање овог проблема. У овом раду је предвиђано да ли ће се код пацијента јавити ово обољење на основу његовог здравственог стања. Обучени следећи модели: логистичка регресија, наивни Бајес, метод потпорних вектора, модел К најближих суседа, метод насумичне шуме, Light Gradient-Boosting Machine, eXtreme Gradient Boosting и Categorical Boosting (CatBoost). Последњи модел се испоставио као најбољи над коришћеним скупом података. Интерпретацијом овог модела дискутовани су фактори који највише утичу на предикцију појаве срчаног удара.*

Кључне ријечи: *срчани удар, машинско учење, CatBoost, XGBoost, SHAP values.*

Abstract – *A heart attack or acute myocardial infarction is the death of part of the heart muscle due to a sudden cessation of circulation through one of the arteries that feed the heart. It is among the most common causes of death in developed and developing countries. It affects men more often than women, and after entering the climax, the risk of the disease becomes even. It is more common in people over 40 years old. Cardiovascular risk factors differ; some can be influenced, and others cannot. The process of detecting or predicting a heart disease in a person can be very unpredictable, long-lasting, and often unsuccessful. Thus, applying artificial intelligence to solve this problem is becoming more common today. This paper used machine learning models to predict whether a patient will have a heart attack based on his health condition. The following models were trained: logistic regression, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest method, Light Gradient-Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost) and Categorical Boosting (CatBoost). The last model was the best-performing model, and it was interpreted to uncover factors most affecting heart attack prediction.*

Keywords: *heart attack, machine learning, CatBoost, XGBoost, SHAP values.*

1. УВОД

Срчани удар, односно акутна миокардијална инфаркција, је једна од манифестација исхемијске болести срца која је најчешћи узрок смрти у земљама широм света. Фактори ризика који предиспонирају одређену популацију да оболи од срчаног удара могу бити разни, али се сматра да међу главне спадају: атеросклероза, старост, пол и повишени крвни притисак. С обзиром на висок степен смртности када је у питању ово обољење, машинско учење је веома корисно јер може помоћи у предикцији ове болести код пацијената на основу њиховог здравственог стања.

У овом раду су тренирани модели машинског учења да, на основу параметара биохемијске анализе и физичког стања пацијената, предвиђају да ли ће се код особе јавити срчани удар или не. Предикција се врши на основу горе поменутих, али и многих других фактора који утичу на појаву овог обољења. Модел са највишим перформансама је интерпретиран у циљу откривања фактора који највише утичу на предикцију срчаног удара.

У наредном поглављу дат је преглед сродних радова. Детаљи решења и имплементације система су описани у поглављу 3, а поглавље 4 садржи анализу добијених резултата. Последње, 5. поглавље, представља закључак рада.

2. ПРЕТХОДНА РЕШЕЊА

Након одабира теме рада, извршена је детаљна анализа сродних истраживања и других радова на исту или сличну тему. На основу резултата тих радова, одлучено је која методологија ће бити примењена као и сам циљ којем се тежи.

У раду [1] илустровани су доступни скупови података везани за срчане болести који су углавном сирове природе, што је сувишно и недоследно. Дакле, постоји потреба за претходном обрадом оваквих

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

скупова података. Скуп високо-димензионалних података мора се свести на скуп нискодимензионалних. Рад [1] указује и на значај издвајања кључних карактеристика из скупа података.

Избор важне карактеристике смањује рад на обучавању алгоритма и самим тим резултира смањењем временске сложености. Време игра и виталну улогу у доказивању ефикасности сличних алгоритама.

Рад [2] се бави имплементацијом модела учења за анализу стварних случајева срчаног удара на основу различитих атрибута. Коришћени модели су логистичка регресија, стабло одлучивања, метод насумичне шуме и модел К најближих суседа, од којих су последња два дала најбоље резултате. Сходно томе, метод насумичне шуме и модел К најближих суседа су имплементирани и у овом раду. Скуп података коришћен у раду [2] садржи 303 инстанце и представља део података искоришћених и за ово истраживање.

Тема рада [3] јесте "Предикција срчаног удара коришћењем машинског учења" и методологија обухвата само логистичку регресију.

Решење овог пројекта евалуирано је уз помоћ унакрсне валидације. Тачност добијена за логистичку регресију јесте 87%. Скуп података коришћен у раду [3] јесте евидентно шири односно обимнији у односу на рад [2], али на основу тачности која је добијена и додатне анализе, закључак је да постоје модели машинског учења који дају боље резултате у овом случају иако је и овај проценат врло висок.

Истраживање рада [4] се првенствено фокусира на скуп података о пацијентима, који се претходно обрађује, а затим се над њим примењују различити алгоритми машинског учења (*KNN, SVM*, логистичка регресија, метод насумичне шуме, наивни Бајес, неуронска мрежа). Коришћени скуп података исти је као и скуп података уврштен у рад [2], с тим што је у раду [4] имплементирана и *SMOTE* (*Synthetic Minority Oversampling Technique)* техника због неуравнотежене природе овог скупа података. Ова техника синтетичког узорковања примењена је и у овом раду као и већина модела машинског учења из рада [4].

Пројекат описан у раду [5] представља упоређивање перформанси различитих класификатора – стабла одлучивања, метода потпорних вектора, метода насумичне шуме и метода наивни Бајес, од којих је већина искоришћена и у овом пројекту. Рад [5] предлаже и класификатор ансамбла који врши хибридну класификацију комбинујући најбоље карактеристике и јаких и слабих класификатора.

Коришћен је *UCI Heart Disease* скуп података који обухвата 76 атрибута који су сведени на 14 најбитнијих. У рад [5] уврштен је и *XGBoost* (*eXtreme Gradient Boosting*) модел машинског учења, који представља један од најбитнијих модела обрађених и у овом раду. Модел предвиђања креиран је након поређења тачности сваке од техника машинског учења обрађених у раду [5]. Циљ је био искористити различите метрике евалуације, као што су матрица конфузије, тачност, прецизност, одзив и *F1*-мера. Класификатор *XGBoost* имао је највећу тачност (81%), када се упореде сви тестирани класификатори.

3. МЕТОД

У наредним поглављима су описани скуп података коришћен у раду и имплементација система за предикцију срчаног удара.

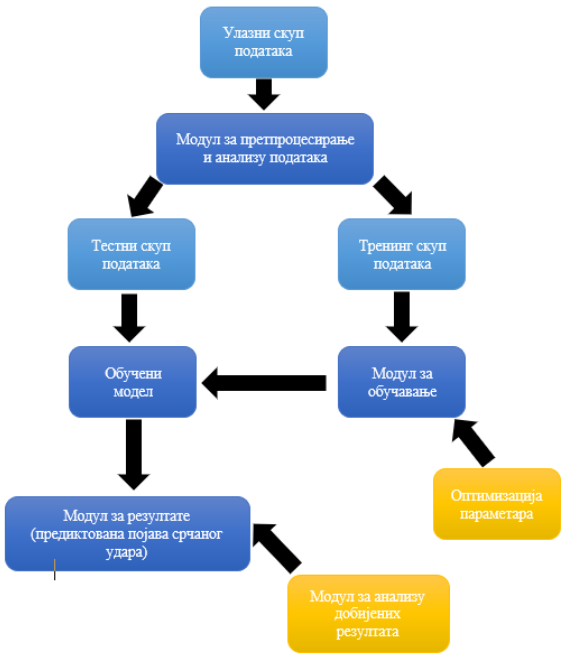
3.1. Скуп података

Скуп података који је коришћен у овом раду обухвата 1190 инстанци и преузет је са линка [6]. Он обједињује четири скупа података који се односе на пацијенте из следећих области: Мађарске, Кливленда, Швајцарске и Лонг Бича. Ови скупови података садрже 14 атрибута, који су, након истраживања и анализе других радова са истом темом, кориговани у 12 најбитнијих: године (*age*), пол (*sex*), тип бола у грудима (*chest_pain_type*), крвни притисак у мировању (*resting_bp_s*), холестерол (*cholesterol*), ниво шећера у крви (*fasting_blood_sugar*), електрокардиографија у мировању (*resting_ecg*), максималан број откуцаја срца (*max_heart_rate*), ангина изазвана вежбањем (*exercise_angina*), ST депресија изазвана вежбањем у односу на одмор (*oldpeak*), максимум ST сегмента изазван вежбањем (*ST_slope*) и циљни атрибут (*target*).

Дакле, овај скуп податак садржи више предикторских (независних) променљивих и једну циљну (зависну) променљиву која представља исход и која је бинарна вредности (0 – срчани удар се није догодио, 1 – срчани удар се догодио).

3.2. Имплементација система за предикцију срчаног удара

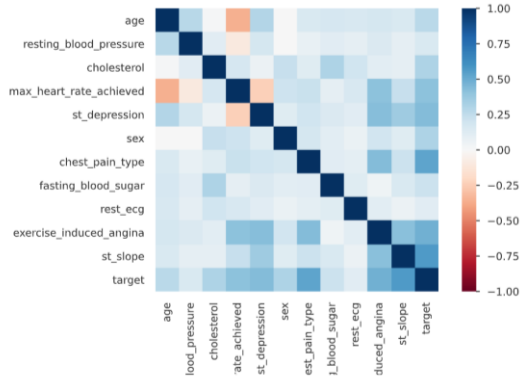
У овом поглављу је представљена имплементација система за предикцију срчаног удара, који се састоји од неколико модула (Слика 1).



СЛИКА 1 - СИСТЕМ ЗА ПРЕДИКЦИЈУ СРЧАНОГ УДАРА

Модули од највећег значаја јесу модул за претпроцесирање и анализу података, модул за обучавање и модул за генерисање резултата.

У оквиру модула за претпроцесирање и анализу података извршена је експлоративна анализа података и извршено је профилисање података. Као резултат профилисања скупа података помоћу *ProfilerReport*-а *Pandas* библиотеке, добијен је детаљнији опис инстанци, одређена корелација између различитих атрибута, као и информација о постојању дупликата и одступања. Утврђено је да улазни скуп података обухвата више мушких него женских пацијената, да је распон година широк – између 28 и 77 година, као и да има највише пацијената са крвним притиском између 100 и 150. Корелација атрибута дата је на Слици 2, са које се може видети да атрибут који директно утиче на појаву срчаног удара јесте тип бола у грудима, као и ангина изазвана вежбањем у корелацији са максимумом ST сегмента изазваног вежбањем. На појаву овог обољења најмање утиче максималан број откуцаја срца у корелацији са годинама пацијента.



СЛИКА 2 - КОРЕЛАЦИЈА АТРИБУТА

Применом експлоративне анализе (*Exploratory Data Analysis, EDA*) закључено је да је већа вероватноћа да се срчани удар јави код мушкараца, да на то највише утиче асимптоматичан тип бола у грудима и повишени притисак. Такође, већа је вероватноћа појаве срчаног удара уколико је вредност атрибута који се односи на максималан број откуцаја срца мања, што и јесте логично – што мањи број откуцаја срца, већи су изгледи да ће се срчани удар догодити.

Пре обучавања модела, скуп података подељен је у тренинг и тест скуп у односу 80:20. Тренинг скуп је улаз у модул за обучавање и над тим подацима је извршена петострука унакрсна валидација за оптимизацију хипер-параметра и примењена *SMOTE* техника. За потребе предикције срчаног удара која је описана у овом раду су изабрани следећи модели: логистичка регресија, наивни Бајес, метод потпорних вектора, модел K најближих суседа, метод насумичне шуме, *LightGBM*, *XGBoost* и *CatBoost*. Након обучавања, модели су примењени на тест скупу података.

Као резултат сваког од поменутих алгоритама машинског учења, у оквиру модула за генерисање резултата израчуната је *F*-мера - мера евалуације која комбинује прецизност и одзив. Одлучено је да би рачунање *Fbeta*-мере, при чему је *beta* параметар већи од 1, било погодније за сам предмет и циљ овог рада, јер би се на тај начин минимизовао број лажно негативних (*FN – false negative*) случајева. Резултати добијени у овом модулу додатно су анализирани уз помоћ *SHAP* вредности и кластеровања.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Сви коришћени модели су тестирани на тест скупу података који је издвојен из првобитног скупа података. Након испробавања више вредности за параметар *beta*, дакле на основу мерења са различитим вредностима овог параметра, одлучено је да се користи *F3*-мера. Ова мера дала је боље резултате у односу на друге испробане *Fbeta*-мере и код ње је вредност лажно негативних случајева максимално минимизована. Вредности ове мере евалуације, као и вредности прецизности и одзива дате су у Табели 1.

Класификатор	Прецизност	Одзив	*F3*-мера
CatBoost	**83.3%**	**99.2%**	**97.4%**
XGBoost	82.8%	99.2%	97.3%
Метод насумичне шуме	75.4%	100%	96.8%
LightGBM	74%	99.2%	95.9%
Метод потпорних вектора	77.8%	97.6%	95.2%
Логистичка регресија	77.8%	97.6%	95.2%
Модел К најближих суседа	73.7%	97.6%	94.5%
Наивни Бајес	65.1%	99.2%	94.3%

ТАБЕЛА 1 - ПРИКАЗ ВРЕДНОСТИ ПАРАМЕТАРА ПРЕЦИЗНОСТ, ОДЗИВ И F3-МЕРА

Оно што можемо уочити из Табеле 1 јесте да су сви имплементирани модели машинског учења високих перформанси. Најбоље резултате дао је *CatBoost*, док наивни Бајесов модел постиже најслабије резултате за предикцију срчаног удара. Највише се истичу модели новије генерације а то су *CatBoost* и *XGBoost* са *F3*-мером респективно 97.4% и 97.3%. Ова два модела по својим перформансама надмашују и све моделе споменуте у сродним радовима који су описани у поглављу 2. Најбољи резултат постигнут је у истраживању [4] и он износи 96.1%. Скуп података коришћен у истраживању [4] је четири пута мањи у однос на скуп података коришћен у овом раду, док су обележја скупа идентична. Остали модели машинског учења истакнути у Табели 1, који су коришћени и у већ поменутим радовима, дали су боље резултате у односу на резултате тих радова. Иако скуп података коришћен у овом поглављу не спада по броју података у велике скупове, ипак је број инстанци које он обухвата већи од броја ентитета у скуповима података већине анализираних постојећих радова.

На Слици 3 приказан су резултати *SHAP* метода примењеног над подацима над којима је обучен *CatBoost* модел. Анализом ових резултата уочено је да, уколико код особе постоји асиптоматичан тип бола у грудима или је вредност *st_slope_flat* атрибута висока, велика је вероватноћа да се јави срчани удар. Притом, холестерол, године и количина шећера у крви значајно доприносе овој појави. Исто тако, особе без срчаних обољења углавном имају вредност 51 *normal* за *st_slope* атрибут и мања је вероватноћа да се срчани удар јави код жена.

Приликом кластеровања коришћена су два најпопуларнија алгоритма, *K-Means* и *DBScan*, и то оба над

SHAP вредностима. Након извршеног кластеровања, генерисана су правила путем *ScopeRules*-а, како би се аутоматски добили одговори који атрибути и њихове вредности су најбитније за који кластер. Узимајући у обзир резултате кластеровања и пронађена правила, може се закључити да су неки од атрибута, који имају највећи утицај на појаву срчаног удара мушки пол, асимптоматичан тип бола у грудима, повишен холестерол и ST сегмент изазван вежбањем са вредношћу *flat*.

СЛИКА 3 - SHAP ВРЕДНОСТИ

У погледу повлачења паралеле између ових резултата и резултата сродних радова, закључује се да се они делимично поклапају. Неки од атрибута, који су се показали као најутицајнији исти су у овом и другим истраживаним радовима описаним у поглављу 2. Са друге стране, постоје и атрибути као што је холестерол, који се анализом *SHAP* вредности и кластеровања истиче као врло битан у овом раду, док га један од радова [3] наводи као потпуно нерелевантан.

5. ЗАКЉУЧАК

У раду је приказан систем за предикцију срчаног удара коришћењем алгоритама машинског учења. Решење овог проблема довело би до знања који то фактори највише утичу на појаву акутног инфаркта миокарда, што би могло помоћи у превенцији таквих случајева.

Доступни подаци су претпроцесирани и анализирани како би се добио детаљнији опис инстанци и установио однос између различитих атрибута. На тренинг скупу података обучено је више модела. Коришћене мере евалуације су *F3*-мера, прецизност и одзив. У овом раду добијена су одређена побољшања тачности у односу на радове који су описани у поглављу 2. Такође, коришћени су новији и напреднији модели машинског учења који су по својим перформансама надмашили све остале. Коришћене су *SHAP* вредности, које на прегледан начин приказују утицај различитих параметара на крајњи исход и које нису коришћене у радовима који су анализирани. Извршено је и кластеровање са извођењем правила за сваки кластер, које даје нове информације о атрибутима и њиховим вредностима.

Један од главних недостатака овог рада представља релативно мали скуп података. Прикупљање додатних узорака, као и укључивање нових атрибута би побољшало резултата рада. Због испробавања већег броја различитих методологија у раду, остаје простор за даљу оптимизацију хипер-параметара модела који су имплементирани, зарад побољшања њихових перформанси. Машинско учење може да буде веома корисно јер може помоћи људима у откривању болести у раној фази на основу њиховог здравственог стања. Због свега наведеног, циљ овог рада и јесте био формирање модела који ће на основу одређених параметара и њихових вредности утврдити да ли постоји вероватноћа да особа добије срчани удар. Узимајући у обзир добијене резултате, машинско учење се показало као поуздано решење за предикцију ове болести. Треба напоменути да је веома битно одабрати одговарајуће атрибуте, класификаторе и методологије које ће се применити. Такође, могуће је и одрадити симулацију коришћењем других алата како бисмо упоредили резултате и видели да ли они доприносе побољшању истих.

6. ЛИТЕРАТУРА

[1] Sultana, M., Haider, A., & Uddin, M. S. (2016, September). Analysis of data mining techniques for heart disease prediction.In 2016 3rd international conference on electrical engineering and information communication technology (ICEEICT) (pp. 1-5). IEEE.

[2] Nabeel, M., Awan, M. J., Raza, M., Muslih-Ud-Din, H., & Majeed, S. (2021, November). Heart Attack Disease Data Analytics and Machine Learning. In 2021 International Conference on Innovative Computing (ICIC) (pp. 1-6). IEEE.

[3] Bhardwaj, A., Kundra, A., Gandhi, B., Kumar, S., Rehalia, A., & Gupta, M. (2019). Prediction of heart attack using machine learning. IITM

[4] Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., & Alharbey, R. (2021). An efficient SMOTE-based deep learning model for heart attack prediction. Scientific Programming, 2021

[5] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[6] Heart-Disease-Dataset https://www.openml.org/search?type=data&status=active&id=43682&sort=runs

Кратка биографија:

Андријана Јеремић рођена је 1997. године у Београду, Република Србија. Основне академске студије завршила је 2019. године на Факултету техничких наука, на ком брани и мастер рад 2023. године из области Примењене рачунарске науке и информатика– Електронско пословање.

контакт: andrijana.jeremi@gmail.com