Предлог пројекта

Овај документ представља предлог пројекта из предмета Системи за истраживање и анализу података. На почетку документа биће објашњена тема и дефиниција пројекта, као и мотивација за његов одабир и рјешавање. Након тога дат је кратак преглед релевантних радова који рјешавају сличне проблеме и користе исти извор података. Идуће што се описује јесте методологија и метод еваулације. Укратко је дефинисан и план рада на пројекту, те наведени софтвери који ће бити кориштени приликом израде пројекта.

Тема пројекта је предикција петогодишњег преживљавања дјеце и адолестената, усљед детекције карцинома – Евинговог саркома. Извор података садржи податке о пацијентима САД-а од 1975.-2017. године, код којих је детектована и праћена ова врста болести.

Дефиниција проблема

Циљ пројекта јесте да се на основу параметара одреди стопа преживњавања пацијента, када је детектован карцином – Евингов сарком. Осим предицкије планирано је да се оптимизују параметри који имају највећи утицај на тачност предикције.

Мотивација

Сваке године око 200 деце и младих у САД-у добије дијагнозу Евинговог тумора, најчешће је то Евингов сарком. Евингов сарком чини 1% свих тумора код деце и адолесцената. На преживљавање унутар првих 5 година утиче много фактора, као што су претходно здравствено стање пацијента, затим стадијум болести у тренутку кад је откривена, величина тумора, операција, хемотерапија, радиотерапија итд. Индивидуална процена прогнозе може бити корисна за саветовање пацијента са раком о избору лечења и за оптимизацију терапијских метода у сарадњи са онколозима и хирурзима.

Релевантна литература

Пронађени су слични радови на ову тему. Тема сличних радова је углавном анализа лечења болести, предикција преживњавања (месец, шест месеци, година, двије године, пет година, десет година итд.) или предвиђање фактора који утичу на преживљавање усљед детекције карцинома. Није пронађен ниједан рад који предвиђа петогодишње преживљавање усљед детекције Евинговог саркома код педијатријских пацијената, а да користи исти скуп података и исту методологију.

Одабрана релеванта литература:
[1] Limin Yang, Tetsuya Takimoto and Junichiro Fujimoto (2014) *Prognostic model for predicting overall survival in children and adolescents with rhabdomyosarcoma ,*
https://link.springer.com/article/10.1186/1471-2407-14-654

Тема рада: Сврха овог рада је била развити прогостички модел преживљавања педијатријских пацијената са Рабдомисаркомом (РМС) користећи параметре који се бележе током стандардног клиничког лечења.
Подаци: Подаци су преузети са националног института за канцер. СЕЕР Програм (*The Surveillance, Epidemiology, end End Results* Program) пружа информације о пацијентима с дијагнозом карцинома.

Фокус је био на идућим атрубутима: старост пацијента, величина тумора, пол пацијента, раса, хистолошки тип, стадијум тумора, операција и радиотерапија.
Коришћени алгоритми: Cox Proportional hazards model је кориштен да би се предвидело петогодишње и десетогодишње преживљавање. За избор модела кориштена је информациона техника Акаике (Akaike information tehnique).
Остварени резултати: Од укупно 1679 пацијената, 543 је умрло. Петогодишња стопа преживљавања била је 65.5% (95% интервал повјерења). Утврђено је да на стопу преживљавања утичу старост пацијента, величина тумора, хистолошки тип, стадијум тумора, операција и радиотерапија.
Закључак: Ова рад има сличан циљ, а то је предикција преживљавања након детекције болести. У овом раду поред предикције петогодишњег преживљавања, предвиђа се и десетогодишње. Користиће се исти извор података, с тим да ће се користити подаци пацијената који имају други облик болести (Евингов сарком). У односу на овај рад користиће се другачија методологија. Из овог рада употребиће се подаци који се односе на битне атрибуте које утичу на предвиђање преживљавања усљед детекције карцинома.

[2] Mogana Darshini Ganggayah, Nur Aishah Taib, Yip Cheng Har, Pietro Lio and Sarinder Kaur Dhillon (2019) *Predicting factors for survival of breast cancer patients using machine learning techniques*, *https://link.springer.com/article/10.1186/s12911-019-0801-4*
Тема рада: Циљ овог истраживања био је предвиђање фактора који утичу на преживљавање пацијената код којих је детектован карцином дојке.
Подаци: Кориштен је скуп података о пацијентима са дијагнозом ове болести, преузет са Медицинског центра универзитета Малезије. Скуп података садржи информације о дијагнози између 1993. и 2016. године. Само једна промјенљива унутар скупа је зависна, а то је статус преживљавања (мртав/жив). Скуп је садржао и 23 независне промјенљиве, које у одређеној мјери утичу на стопу преживљавања.
Коришћени алгоритми: Квалитет података упоређен је помоћу шест алгоритама: Decision Tree, Random Forest, Neural Networks, Extreme boots, Logistic Regression, Support Vector Machine. Скуп података је подељен у тренинг и тест податке у омеру 70:30%. Сви алгоритми користили су исти омер. Сваки модел је оцењен тачношћу, осетљивошћу, специфичношћу, прецизношћу, коефицијеном корелације, кривом прецизности и опозива и калибрационом кривом.
Остварени резултати: У погледу тачности модела и мере калибрације, сви алгоритми су дали блиске резултате. Најмања тачност је постигнута методом стабла одлучивања (79.8%), а највећа методом RF (82.7%). Најважније променљиве за ово истраживање су класификација стадијума карцинома, величина тумора, број уклоњених аксиларних лимфних чворова, врсте примарног лечења и методе дијагнозе.
Закључак: Путем овог рада добија се јаснија слика о атрибутима битним приликом предикције преживљавања пацијената обољелих од рака. Осим тога, у релевантном раду кориштена је методологија, слична планираној за израду овог рада.

[3] Leili Tapak, Nasrin Shirmohammadi-Khorram, Payam Amini, Behnaz Alafci, Omid Hamidi, Jalal Poorojal (2019) *Prediction of survival and metastasis in breast cancer patients using machine learning classifiers*, *https://www.ceghonline.com/article/S2213-3984(18)30182-9/fulltext*

Тема рада: Предвиђање преживљавања и метастаза код пацијената са раком дојке помоћу класификатора машинског учења. Сврха овог рада је упоређевање перформанси шест техника

машинског учења и две традиционалне методе за предвиђање преживљавања рака дојке и метастаза.
Подаци: Кориштен је скуп података који потиче из кохортног истраживања, спроведеног 2014. године у Техерану, Иран. Анализиране су информације о пацијентима са дијагнозом рака дојке од 1998. до 2013. године. Подаци су регистровани у Свеобухватном центру за контролу рака (Comprehensive Cancer Control Center). Атрибути на којима је био фокус су: статус преживљавања (жив/мртав), постојање метастаза (да/не), старост пацијента, степен болести, стадијум болести, рецептор за естроген (позитиван/негативан), рецептор за прогестерон (позитиван/негативан), хирушки приступ.
Кориштена је стратегија унакрсне валидација и у предвиђању преживљавања и у предвиђању метастазе пацијента према два сценарија. Први свенарио је подела на транинг и тест податке у односу 70:30%, а други сценарио је подела тренинг/тест у односу 50:50%.
Коришћени алгоритми: Naive Bayes (NB), Random Forest (RF), AdaBoost, Support Vector Machine (SVM), Least-square SVM (LSSVM), Adabag,
Logistic Regression (LR) и Linear Discriminant Analysis.
Остварени резултати: У предвиђању преживљавања просечна спрецифичност свих техника била је ≥ 94%. SVM i LDA имају већу осетљивост (73%) у односу на остале технике. Већу укупну тачност (93%) имају алгоритми SVM и LDA. Приликом предвиђања метастаза RF је имао највећу специфичност (98%), а NB највећу осјетљивост (36%), а највећу укупну тачност LR и LDA (86%). Крајњи закључак је да алгоритам SVM надмашује све остале методе машинског учења у предвиђању преживљавања пацијената у неколико критеријума, такође и LDA је показао сличне перформансе.
Закључак: У овом раду се осим предвиђања преживљавања, предвиђају и метастазе. Рад се сматра релевантном литературом, јер користи методологију која се једним дијелом подудара са планираном методологијом предвиђања преживљавања усљед детекције Евинговог саркома код деце и адолесцената.

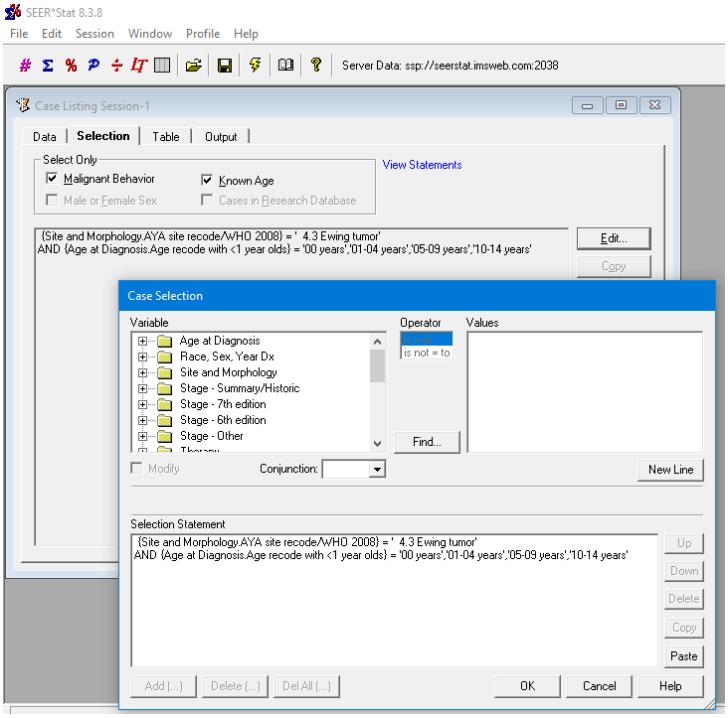
Скуп података

Да би се предвидела петогодишња стопа преживљавања усљед детекције Евинговог саркома користиће се CEEP програм, односно подаци пацијената са националног института за рак. Подацима се приступа путем идуће веб странице: *https://seer.cancer.gov/data/access.html .* Прије приступа подацима неопходно је да се попуни форма са личним подацима, као и намени кориштења података о пацијентима. Осим тога, неопходно је потписати и прослиједити уговор о коришћењу података о пацијентима. Након успјешне пријаве, подацима се приступа путем SEER*Stat десктоп апликације, доступне на сајту: *https://seer.cancer.gov/seerstat/software/*.

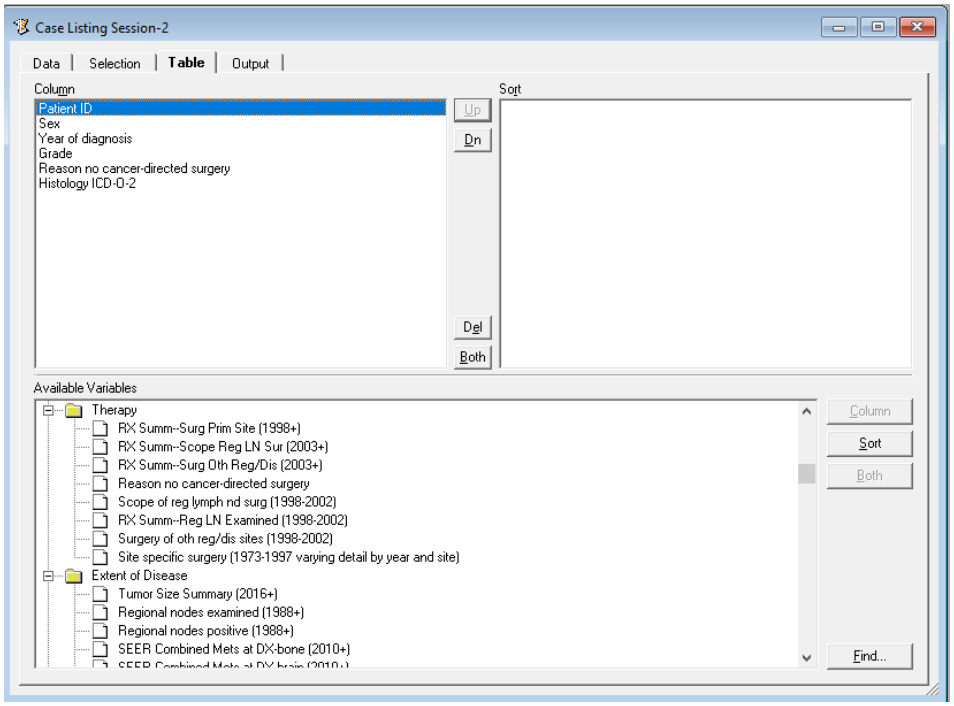
Скуп података обухавта податке о пацијентима из девет регистара CEEP-а. Код пацијената је дијагностикован и праћен карцином од 1975. – 2017. године. Подаци се добијају задавањем упита унутар регистра. Упит се задаје интерактивно. Приликом задавања упита наводе се жељени атрибути. Добијени резултати се експортују у .csv fajl. На слици је приказан изглед упита над регистрима, гдје су селектовани само пацијенти са детектованом болести Евингов сарком, који имају између нула и петнаест година. Након селекције пацијената наводи се који подаци (атрибути) да се прикажу за селектоване пацијенте. Због великог броја атрибута, као и широке и нове области изучавања, сви постојећи нису до сада истражени. Потребно је извршити детаљну анализу података у регистрима. Јавља се проблем недостајућих вриједности. За поједине атрибуте недостају вриједности (нпр. забиљежени су подаци о том атрибуту тек од 2016 .године) . Примјер одабира атрибута дат је на слици број 2. На слици су изабрани атрибути: ИД, спол, година

детектовања, стадијум… док су испод дате могућности додавања нових атрибута.

У девет регистара налазе се подаци о 1942 пацијената, код којих је детектован Евингов сарком, од којих 1068 пацијената спада у групу од 0-15 година – деца и адолесценти.



Слика 1. Приказ креирања упита приликом ког се селектују само одређени пацијенти



Слика 2. Интерактивни одабир атрибута унутар апликације

Методологија

Кораци методологије:

- Припрема података – Из поменутог скупа података, потребно је издвојити само оне пацијенте који задовољавају старосни критеријум (децу и адолесценте) и оне код којих је дијагностификован Евингов сарком. Неопходно је пронаћи атрибуте који највише утичу на процену преживљавања. Процес процене атрибута је проширење овох проблема. Користиће се методологија за процену: Decision Tree, Random Forest, Neural Networks, Extreme boots, Logistic Regression, Support Vector Machine.
- Добијене податке потребно је анализирати и обрадити.
- Алгоритми за класификацију: Random Forest, Extreme boots, Logistic Regression, Support Vector Machine, Naïve Bayes. *Избор алгоритама ће зависити од одабира података и могуће су промене у избору током израде пројекта.*
- Високе димензије и мала количина узорака у великој мери повећавају тежину анализе преживњавања. Из тих разлога, за предикцију, користиће се модел полу-надгледаног учења: Полу-нагледана логистичка регресија и Кластеровање.
- Тестирање модела над подацима
- Поређење резултата различитих алгоритама

Метод евалуације

Узимајући у обзир ограничен број пацијената у овој области истраживања, користиће се стратегија унакрсне валидације на тренинг скупу.

Софтвер

Подацима се приступа помоћу *SEER*Stat* апликације. Data Mininig анализе ће бити изведене уз помоћ софтвера за обраду и анализу података *Rapid Miner*-а. У случају да алат нема све потребне функционалности, користиће се библиотеке за анализу података у програмском језику *Python*. Такође, апликација ће бити израђена у *Python* програмском језику.

План

План рада обухвата идуће фазе:

- Прикупљање података
- Анализа и обрада података
- Креирање модела
- Верификација модела
- Визуализација резултата
- Анализа и упоређивање добијених резултата

Тим

Тања Станић E2 47/2020

Сара Челик