Математичка гимназија

МАТУРСКИ РАД

Програмирање и програмски језици

Машинско учење и проблем откривања релација између лекова и узрочника болести

Ученик:
Михаило Милошевић IVд

Ментор:
Јелена Хаџи-Пурић

Београд, јун 2020.

Садржај

1 Увод **1**

1.1 Проблеми фармацеутске индустрије 1

1.2 Процес откривања лека 2

1.2.1 Идентификација узрочника болести 2

1.2.2 Преклиничко сужавање избора 2

1.2.3 Преклиничка медицинска хемија 3

1.2.4 Преклиничка *In Vitro* изучавања 3

1.2.5 *In Vivo* изучавања на животињама 3

1.2.6 Клиничке провере 4

1.2.7 Комерцијализација 4

1.3 Биоинформатика 5

2 Методологија **6**

2.1 Алгоритми машинског учења 6

2.2 Надгледано учење 7

2.2.1 Алгоритам K најближих комшија (KNN algorithm) . 7

2.2.2 Стабло одлуке (Decision Tree) 8

2.2.3 Алгоритам векторски потпомогнуте машине (SVM algorithm) 9

3 Припрема података за анализу **11**

3.1 DrugBank 11

3.2 Конверзија података из .sdf у .csv датотеку 12

3.3 Објашњење начина припреме података 13

4 Обрада података и учење машинског модела **14**

4.1 Обрада података 14

4.2 Процес обраде података 14

5 Резултати и дискусија **16**

6 Закључак **18**

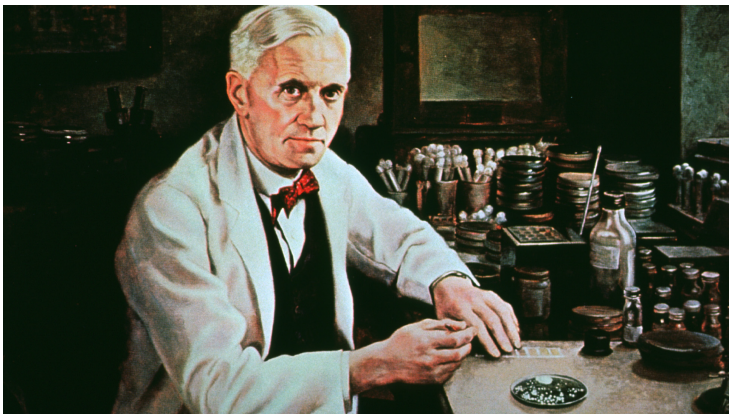
7 Референце **20**

1

Увод

1.1 Проблеми фармацеутске индустрије

Живот у данашње време постаје све бржи. Због оваквих промена почињу да се развијају комплекснији и отпорнији паразити, који код људи изазивају теже и захтевније болести. Сходно томе ако се научници предају традиционалним начинима отривања лекова, у неком тренутку понестаје им оног најбитнијег фактора - времена. У данашњим ситуацијама један дан може да буде разлика између спашеног и изгубљеног живота.



Слика 1.1: Александар Фелминг у својој лабораторији

Најпознатији пример открића неког лека, који је изазвао велике промене у фармацеутској индустрији, био је пример Александра Флемнига.

Овај научник је 1928. својом несмотреношћу оставио узорак на коме је радио, у отвореној Петри шољи, поред свог прозора док је био на једномесечном путу. Након повратка изоловао је новонасталу супстанцу и сасвим случајно открио лек који је данас у широкој употреби као антибиотик - Пеницилин.

Оваква открића су веома значајна, али фактор среће није увек присутан. Научници уз напоран рад проверавају многобројне супстанце у ситуацијама избијања нове болести и пореде њихове учинке. Процес откривања истих је стандардан и до сада је давао резултате, али кроз дуготрајне временске периоде. ([1])

1.2 Процес откривања лека

Иако је технологија напредовала на свим пољима, у фармацеутској индустрији постоји пуно простора за унапређивање. На промени процеса откривања лекова је тешко радити, а да се не угрози сигурност људске популације. Поред тога оваква истраживања су скупа и трају некада и преко 10 година.

Због оваквих карактеристика у данашње време не постоји лек за 90% ретких болести. Да би се оваква ситуација оправдала неопходно је у потпуности имати увид у овај процес. ([1])

1.2.1 Идентификација узрочника болести

Дужина: 2+ година

Први корак откривања лека је идентификација узрочника болести у самом телу јединке. Код људи то су обично мутиране секвенце ДНК ланаца, погрешно синтетисани протеини или неки други биомаркери.

У овом стадијуму постоји на десетине хиљда могућих биомаркера или мутација на ДНК које се морају проверити и зато овај посао изускује доста времена. ([1])

1.2.2 Преклиничко сужавање избора

Дужина: 1-2+ година

Други корак откривања новог лека је провера интеракције многобројних супстанци које би могле реаговати са наведеним узрочницима болести из првог корака. Циљ овог корака је да се значајно смањи број супстанци које потенцијално могу бити лек.

Као што се већ може наслутити, од првог корака зависи и дужина другог корака. Уколико се избор смањи на мањи број узрочника болести, научници ће имати мање комбинација за проверу. ([1])

1.2.3 Преклиничка медицинска хемија

Дужина: 1-2+ година

Трећи корак је детаљнија анализа супстанци које су се показале прикладним у другом кораку. У овој фази се у сврху сужавања избора узима 3D структура молекула и његово пријањање на узрочнике болести. ([1])

1.2.4 Преклиничка *In Vitro* изучавања

Дужина: 1-2+ година

Овај корак је први сусрет могућих лекова са ћелијом. Врше се истраживања у петри шољама тако што се *In Vitro* убацује супстанца и проверава се како ће она утицати на остатак ћелије, али пре свега на узрочника болести. ([1])

1.2.5 *In Vivo* изучавања на животињама

Дужина: 1-2+ година

У петом кораку се одбацује највећи проценат могућих супстанци. Истраживања се обично врше на мишевима у лабораторијама за чији рад је потребно много новца. Уколико се утврди да супстанца није погодна у овом кораку штете су огромне и по буџет научника, али и по живу средину.

Овакав проценат неуспешности се може приписати разлици између изучавања *In Vitro* и *In Vivo*. Иако *In Vitro* истраживања дају добре резултате, организми који конзумирају дате супстанце су много комплекснији и тешко је предвидети исход. ([1])

1.2.6 Клиничке провере

Дужина: 6+ година

Након успешности у свим претходним корацима супстанце су спремне за проверу сигурности и ефикасности на људском систему. Овај корак се састоји из три подкорака од којих сваки посебно може коштати и преко 20 милиона америчких долара. ([1])

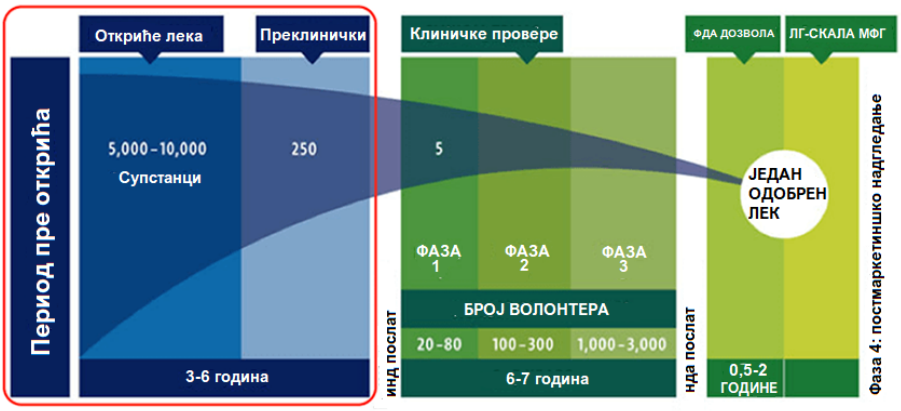
1.2.7 Комерцијализација

Дужина: 1+ година

Након свих ових корака и након одобрења од стране светске здравствене организације, лек може бити пуштен у производњу и може се продавати.

Почетне цене новодобијених лекова су назамисливе, али оне су такве због потребе надокнаде утрошених средстава у истраживању.

Након упознавања са процесом и његовом укупном дужином трајања поставља се питање да ли је процес могуће скратити и убрзати. ([1])



Слика 1.2: Приказ процеса откривања лекова

1.3 Биоинформатика

Биоинформатика је интердисциплинарна област која се бави прављењем софтверских алата за разумевање биолошких података, поготово када су они комплексни и велики. Ова област се служи разним алгоритмима обраде великих скупова података, али један од тренутно најактуалнијих је машинско учење. ([4])

Машинско учење (МУ) је научна дисциплина која се бави алгоритмима и статистичким подацима на начин да се компјутерском систему пружају потребне информације да извршава задатке без коришћења експлицитног алгоритма. Ова дисциплина може помоћи у тренутном проблему фармацеутске индустрије. Предност овог процеса је што може имати примену у сваком кораку откривања лека и значајно их може скратити и убрзати откриће новог лека. ([1],[5])

У овом раду фокус ће бити на одређивању успешности предвиђања постојаности релације између одређеног лека и одређеног узрочника болести на другачији начин од свакодневних радова, у циљу проверавања да ли је ниво комплексности програма који се примењује потребан или се може поједноставити. Овај процес се састоји из 3 дела:

- Припреме података за анализу;
- Обрада података и учење машинског модела;
- Одређивање постојања релације према машинском моделу.

2

Методологија

2.1 Алгоритми машинског учења

Машинско учење се дели у 3 категорије:

- Надгледано учење (Supervised learning);
- Ненадгледано учење (Unsupervised learning);
- Појачано учење (Reinforcement learning).

Ненадгледано учење се претежно користи код уједначеног броја припадник сваке класе одговора (у нашем случају „тачно-нетачно“), због његовог начина учења (прво се одређују класе, па се онда проверава припадност класама и коректност). Имајући у виду да је број непостојећих веза за неки узрочник болести и лекове много већи од броја лекова који заправо могу помоћи код лечења посматраног узрочника, овај начин учења не задовољава критеријуме и не би могао са задовољавајућом сигурношћу да раздвоји „добре“ од „лоших“ лекова. ([2],[3])

Појачано учење се користи код проблема који имају велику количину комуникације са корисницима и од којих проблем може добијати повратне информације и самим тим се усавршавати. Овакво учење у овом проблему нема будућност, због тога што је време прикупљања података за овај модел, док он не постане довољно валидан, енормно и због тога што тражимо решење као листу могућих лекова, а не лекове које смо испитали лабораторијски и са сигурношћу знамо њихово дејство. Циљ је

коришћење економичнијег и бржег процеса и да се сузи избор енормног проверавања новосинтетисаних лекова.([2],[3])

Елиминација претходна два начина учења доводи до јединог начина учења које ће бити од користи у овој врсти проблема.

2.2 Надгледано учење

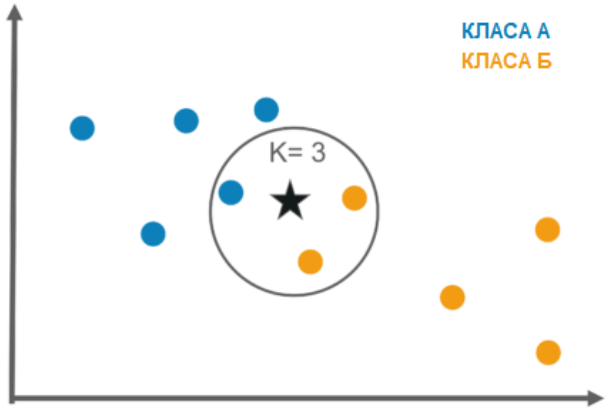
Надгледано машинско учење испуњава највећи број услова, јер се током његовог рада подаци раздвајају у групе према њиховим ознакама,а не само према карактеристикама. Постоје многи алгоритми надгледаног учења, али они који могу послужити за решавање нашег проблема су алгоритми класификације.

Класификација је процес раздвајања података у класе одговора. Решење ових алгоритама је класа којој одређени улаз припада. За сврху овог рада користе се 3 најприкладнија алгоритма класификације:

- Алгоритам К најближих комшија(KNN algorithm);
- Стабло одлуке (Decision Tree);
- Алгоритам векторски потпомогнуте машине(SVM algorithm).

2.2.1 Алгоритам К најближих комшија (KNN algorithm)

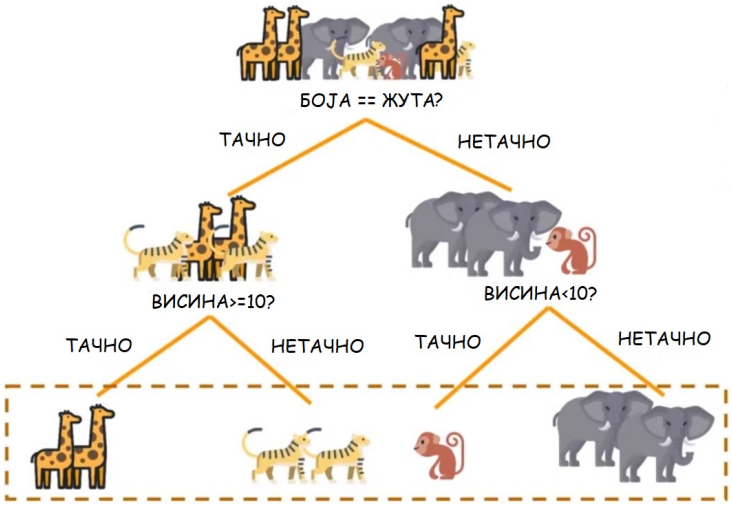
Алгоритам К најближих комшија је најједноставнији алгоритам машинског учења. Заснива се на проналажењу К најсличнијих лекова и одређивања постојаности везе методом бројности. У зависности од бројности одговора сваке класе од К одабраних одређује се одговор за улазни упит. Препоруке многих радова је да се за К узима $sqrt(n)$ где је n број тренинг података. У овом случају то не би било од помоћи, јер је број непостојећих веза реда 10^3, док је број веза неког узрочника болести са лековима реда 10^1. Због наведеног проблема, у овом случају за К се узима вредност 1, који се показао као најоптималнији и најбољи. ([2],[3])



Слика 2.1: Илустрација алгоритма K најближих комшија

2.2.2 Стабло одлуке (Decision Tree)

Стабло одлуке је алгоритам машинског учења који је поред KNN алгоритма најинтуитивнији. Он ради по принципу поделе података на две групе тако да се смањује нека предефинисана вредност. У највећем броју случајева за ту вредност се узима ентропија, мера неуређености система.



Слика 2.2: Илустрација стабла одлуке

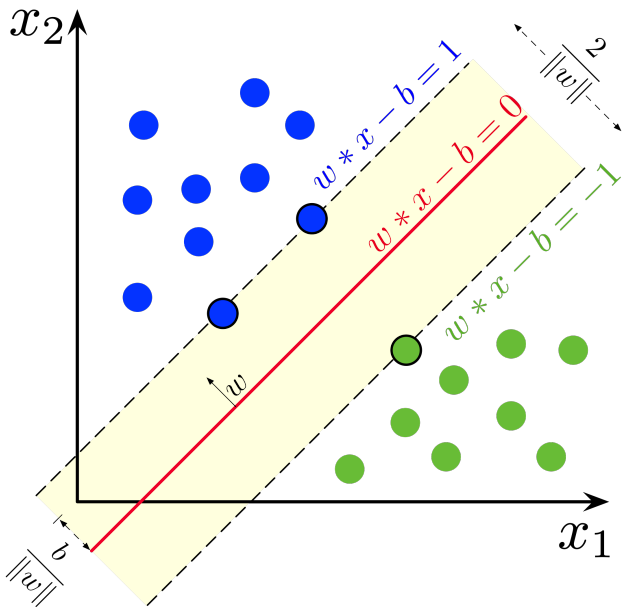
$$S = \sum_{i=1}^{N} P(k_i) log_2(P(k_i))$$

S - Ентропија
N - Број различитих елемената скупа
$P(k_i)$ - Вероватноћа да узмемо i-ти елемент

Дакле, као што и сама дефиниција каже тежимо да направимо бинарно стабло у којем листови садрже минималну ентропију тј. ентропију која је једнака 0. У сврху нашег проблема као класе за разврставање користимо класу „YES“ и класу „NO“. ([3])

2.2.3 Алгоритам векторски потпомогнуте машине (SVM algorithm)

Алгоритам векторски потпомогнуте машине је алгоритам надгледаног машинског учења у коме су подаци представљени као тачке у хипер простору. Овај алгоритам раздваја тај скуп података на класе уз помоћ хиперравни, као такав је један од најкориснијих алгоритама у биоинформатици и широко је у примени у тој области.



Слика 2.3: Илустрација линеарног SVM алгоритма

За најједноставнији облик, линеарне векторски потпомогнуте машине циљ овог алгоритма је да се пронађе права која је еквидистантна од обе класе података. ([2],[3])

$$\vec{w} * \vec{x} - b = 0$$

w - Вектор нормале задате праве
x - Вектор положаја података
b - Слободни члан праве

3

Припрема података за анализу

3.1 DrugBank

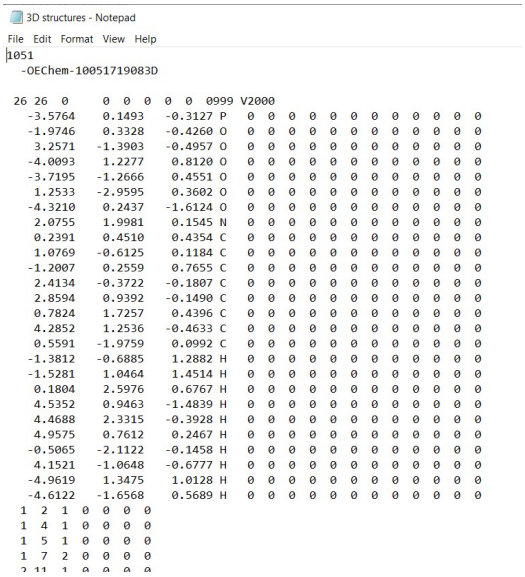
DrugBank база података је први пут пуштена у јавност 2006. године. Ова база се од свог оснивања бави пружањем података о лековима и узрочницима болести. У њеним подацима се налазе детаљне информације структура и карактеристика велике количине лекова. Такође, поред ових информација садржи и информације о релацијама између постојећих лекова и узрочника болести. Због саме структуре и количине података ова база података пружа идеалну подлогу за прављење модела за машинско учење. ([6])

Слика 3.1: Лого DrugBank базе података

Подаци у овој бази података се налазе у разним форматима и због тога се наилази на нови проблем, а то је конверзија. Потреба за конверзијом се јавља због лакшег управљања и учитавања у програмском језику *Python* преко библиотеке *Pandas*.

3.2 Конверзија података из .sdf у .csv датотеку

Екстракција потребних података урађена је уз помоћ програмског језика *c++*. Неки део података који добијамо у нашем оригиналном фајлу није потребан и врши се редукција атрибута уз уклањање вишка информација. Формат .csv је стандардни формат за истраживања у области машинског учења, због чега је ова конверзија била неопходна.



Слика 3.2: Структуре у .sdf формату

<DATABAS	<DATABAS	<MOLECU	<JCHEM_A	<JCHEM_A	<JCHEM_C	<JCHEM_F	<JCHEM_L	<JCHEM_N	<JCHEM_F	<JCHEM_F	<JCHEM_F	<JCHEM_F	<JCHEM_F	<JCHEM_F	<JCHEM_F	<DRUGBANK_ID>
DB00114	drugbank	247.1419	6	20.89889	3	0	-2.08561	1	-2	6.691277	1.680299	4.109012	116.95	54.7463	4	DB00114
DB00116	drugbank	445.4292	12	42.95313	8	0	-4.21263	3	-2	3.863701	3.509569	3.584683	207.27	121.3866	9	DB00116
DB00117	drugbank	155.1546	4	14.67026	3	0	-3.61602	1	0	12.94117	1.849973	9.437137	92	38.059	3	DB00117
DB00118	drugbank	398.44	10	39.26098	4	0	-5.31868	3	1	12.44425	1.70354	9.414366	185.46	107.072	7	DB00118
DB00119	drugbank	88.0621	3	7.310506	1	0	0.065874	0	-1		2.930123	-9.58317	54.37	17.9897	1	DB00119
DB00120	drugbank	165.1891	3	17.03198	2	0	-1.18444	1	0		2.46923	9.446936	63.32	45.1163	3	DB00120
DB00121	drugbank	244.311	3	24.91924	3	0	0.319424	2	-1	13.54699	4.404317	-1.86364	78.43	60.0509	5	DB00121
DB00122	drugbank	104.1708	1	12.56561	1	1	-4.66227	0	1		13.96871	-3.24988	20.23	42.194	2	DB00122
DB00123	drugbank	146.1876	4	15.84375	3	0	-3.21453	0	1		2.738286	10.28806	89.34	37.8095	5	DB00123
DB00125	drugbank	174.201	6	17.7984	5	0	-3.15594	0	1		2.411838	12.411	125.22	53.9231	5	DB00125
DB00126	drugbank	176.1241	5	14.93253	4	0	-1.91356	1	-1	11.1977	4.359917	-2.97973	107.22	37.0321	2	DB00126
DB00127	drugbank	202.3402	4	26.69469	4	0	-1.45401	0	4		11.09925		76.1	62.5582	11	DB00127
DB00128	drugbank	133.1027	5	11.28423	3	0	-3.50375	0	-1	5.11375	1.702544	9.614154	100.62	26.5327	3	DB00128
DB00129	drugbank	132.161	4	13.85326	3	0	-3.65829	0	1		2.666423	10.28682	89.34	33.2085	4	DB00129
DB00130	drugbank	146.1445	4	13.86828	3	0	-4.00113	0	0	16.24489	2.147302	9.312091	106.41	33.1099	4	DB00130
DB00131	drugbank	347.2212	10	29.99861	5	0	-4.78829	3	-2	6.281814	1.226572	4.96539	186.07	74.0685	4	DB00131
DB00132	drugbank	278.4296	2	34.98036	1	0	6.059955	0	-1		4.988167		37.3	89.6354	13	DB00132
DB00133	drugbank	105.0926	4	9.386632	3	0	-3.88774	0	0	15.16607	2.03399	8.931304	83.55	22.041	2	DB00133

Слика 3.3: Испис програма за конверзију

3.3 Објашњење начина припреме података

Да би се машинско учење могло извршити, потребан је критеријум за упоређивање података. Теорија мере, одређивања критеријума, доста црпи научне резултате из области математичке анализе и вероватноће. У претходно објављеним радовима, критеријуми за сличност молекула се своде на алгоритам за проналажење сличности између два тежинска графа. Циљ овог рада је да се провери да ли критеријум може да се постави да буду карактеристике самих молекула, као што су: Ph раствора, поларизибилност, рефрактивност... ([2],[7])

Као што је наглашено у уводном делу, време је некада од пресудног значаја, па ако се једноставнији алгоритам може примењивати са упоредивом прецизношћу и корисношћу, то би убрзало процес проналажења лека.

4

Обрада података и учење машинског модела

4.1 Обрада података

Обрада података подразумева раздвајање потребних информација за учење модела и дељење добијених информација на тест и тренинг податке. Целокупни програми су писани у *Python* развојном окружењу *JupyterNotebook*-у. Сви кодови исписани у сврху извршења овог матурског рада се налазе на следећем линку: https://github.com/ravenstorm2001/Maturski

4.2 Процес обраде података

За потребе коришћења алгоритма првенствено учитавамо податке из конвертованог фајла, а потом их је потребно и дорадити. Неки од података у самој бази нису постојали и на тим пољима у нашем програму је уписана вредност NaN која није погодна. Због тога се та вредност замењује са средњом вредношћу свих постојећих вредности и избегава се проблем недостатка података. Овом начину допуне података се прибегава због повећане мере одзива испитане у претходним радовима. ([3])

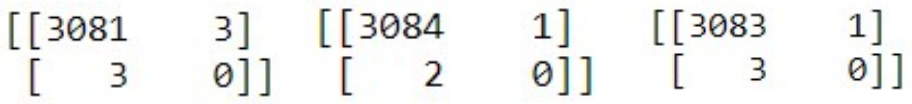
Након успешног учитавања података о лековима учитавају се информације о везама узрочника болести и лекова. Из те датотеке се компонује листа y са одговорима у облику: 1 ако лек реагује са узрочником и 0 ако не реагује.

На крају се овако изоловани подаци раздвајају на тест и тренинг групе и врши се учење одређеног модела машинског учења који је коришћен.

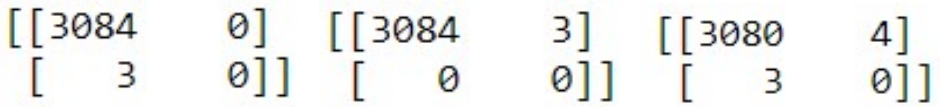
5

Резултати и дискусија

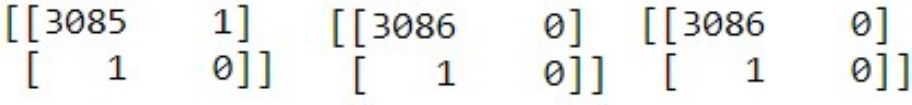
Исписи матрица грешке дају најпрецизнију анализу самог алгоритма и све потребне информације за проверу валидности истог.



Слика 5.1: Приказ решења за Replicase polyprotein 1ab а)KNN; б)Decision Tree; в)SVM



Слика 5.2: Приказ решења за Human immunodeficiency virus type 1 protease а)KNN; б)Decision Tree; в)SVM



Слика 5.3: Приказ решења за Sphingosine 1-phosphate receptor 1 а)KNN; б)Decision Tree; в)SVM

Сви испитивани узрочници болести се верује да имају директну везу са новонасталом болешћу COVID-19.

Пракса је да се код оцењивања ваљаности алгоритма гледају прецизност и пре свега f1 резултат. У овом проблему те вредности нису релевантне, јер се трага за грешкама алгоритма тј. за нетачним позитивним реакцијама. Ако се јави да модел предвиди да неки лек реагује са одређеним узрочником болести, тај лек представља потенцијални лек. Циљ алгоритма је да се пронађе листа могућих лекова за које се до тада веровало да не реагују са одређеним узрочницима болести који су експериментално утврђени. ([3])

Јасно се види да у неким случајевима алгоритам не враћа ниједан нетачан позитиван одговор, што доказује да овај алгоритам у неким ситуацијама тежи преученом стању. То је стање када мале варијације од одређеног податка из неке класе могу дати другачије одговоре.

6

Закључак

Биоинформатика је постала популарна последњих деценија, што због развоја науке, што због развоја рачунара и нумеричких израчунавања. Ови методи машинског учења који за критеријум узимају карактеристике које нису структурне него скаларне, показали су се као валидни. Ови модели су одредили неколико лекова по сваком узрочнику болести и због тога се показују као корисни. Резултати метода који користе структуру као критеријум предвиђају више могућих лекова по узрочнику болести, али то није увек најкорисније. Некада те листе могу бити предугачке и могу одузети превише времена. Мана начина приступа у овом раду може бити да није довољан број откривених лекова и да се не пронађе лек, што би се завршило са неуспехом и потрошеним временом и ресурсима. У бидућности, овај метод приступа би могао да се помоћу тзв. хибридног приступа искомбинује са методом приступа структурама не би ли се формирао алгоритам који ће бити најоптималнији и најекономичнији. ([2])

Захвалница

Посебно се захваљујем Јелени Хаџи-Пурић, професору Математичког факултета Универзитета у Београду, на одвојеном времену, предложеној литератури и на свим коментарима и сугестијама. Захваљујем се и *DrugBank* организацији на доступности инфромација и на одобравању коришћења истих. Такође бих волео да се захвалим *Simplilearn Youtube* страници на корисним предавањима за савлађивање основа машинског

учења у програмском језику *Python*.

7

Референце

1. Joey Mach, 2019, *Unlocking Drug Discovery With Machine Learning*, Medium - Towards Data Science, последњи пут погледан 7 маја 2020, <https://towardsdatascience.com/unlocking-drug-discovery-through - machine-learning-part-1-8b2a64333e07>

2. Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, Shanfeng Zhu,2013, *Similarity-based machine learning methods for predicting drug-target interactions: a brief review*, Oxford University Press, VOL 15. NO 5. 734-747

3. Simplilearn 2019, *Machine Learning Full Course | Learn Machine Learning | Machine Learning Tutorial | Simplilearn*, Simplilearn, последњи пут погледан 7 маја 2020, <https://www.youtube.com/watch?v=9f-Garc DY58&t=12931s>, последњи пут погледан 7 маја 2020.

4. 'Bioinformatics' (2020) Wikipedia, доступан на https://en.wikipedia.org /wiki/Bioinformatics, последњи пут погледан 7 маја 2020.

5. 'Machine Learning' (2020) Wikipedia, доступан на https://en.wikipedia .org/wiki/Machine_learning, последњи пут прегледан 7 маја 2020.

6. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.

7. Hattori M, Okuno Y, Goto S, et al. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. JAmChemSoc 2003;125(39):11853–65.