Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Sandi Mikuš

Ekstrakcija medicinskega znanja iz tekstovnih opisov pri napovedovanju okužbe z rezistentno bakterijo

magistrsko delo

magistrski program druge stopnje
računalništvo in informatika

Mentor: izred. prof. dr. Matjaž Kukar
Somentor: prof. dr. Bojana Beovič, dr. med., višja svetnica

Ljubljana, 2018

Zahvala

Zahvaljujem se družini, ker mi je omogočila študij in me skozi vsa leta šolanja in študiranja podpirala. Zahvaljujem se tudi mentorjema za strokovno pomoč, ter vsem prijateljem, ki so mi pomagali, me podpirali in me zabavali.

Sandi Mikuš, 2018

"No problem can be solved from the same level of consciousness that created it."

— Albert Einstein

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
NLP	natural language processing	obdelava naravnega jezika
NLTK	natural language toolkit	orodje za obdelavo naravnega jezika
ESBL	extended spectrum beta-lactamase	laktamaze beta razširjenega spektra
SGD	stochastic gradient descent	statistični gradientni sestop
SVM	support vector machine	metoda podpornih vektorjev
CA	classification accuracy	klasifikacijska točnost
ROC	receiver operating characteristic	karakteristika delovanja sprejemnika
L-BFGS-B	limited memory	Broyden-Fletcher-Goldfarb-Shanno
	Broyden-Fletcher-Goldfarb-Shanno	algoritem za omejeno
	algorithm for bound-constrained	optimizacijo z
	optimization	omejenim pomnilnikom
AUC	area under curve	ploščina pod krivuljo
DDST	double disk sinergy test	sinergijski test z dvema diskoma
CRISP-DM	cross-industry standard	standardiziran postopek
	process for data mining	za podatkovno rudarjenje
ReLDI	Regional Linguistic Data Initiative	Regionalna iniciativa za jezikovne podatke
WHO	World Health Organization	Svetovna zdravstvena organizacija
ANOVA	analysis of variance	analiza variance

Povzetek

Naslov: Ekstrakcija medicinskega znanja iz tekstovnih opisov pri napovedovanju okužbe z rezistentno bakterijo

Problem odpornosti mikroorganizmov proti protimikrobnim sredstvom, je iz dneva v dan vse večji. Število mikroorganizmov, odpornih proti protimikrobnim sredstvom narašča hitreje od števila novih protimikrobnih sredstev. Če se pri bolniku izbere napačno protimikrobno sredstvo (zdravilo), se lahko odpornost mikroorganizmov še poveča. Mikroorganizmov odpornih proti protimikrobnim sredstvom je kar nekaj, vendar smo se v našem delu osredotočili izključno na primer *Escherichie coli*, ki izloča encime ESBL. Bolniki začnejo prejemati ustrezno protimikrobno sredstvo šele, ko prispejo mikrobiološki izvidi (po približno dveh dneh). Na podlagi zdravnikovega izvida v naravnem jeziku (sestavljenega iz anamneze in statusa pacienta) smo poskušali napovedati, ali je bolnik okužen s prej omenjeno bakterijo *E. coli* ali ne. Čeprav smo dosegli visoko specifičnost (90% za SVM oz. 86% za naivni Bayesov klasifikator), klasifikator zaradi prenizke senzitivnosti (28% za SVM oz. 33% za naivni Bayesov klasifikator) in občutljivosti problema ne more biti uporabljen. Problem vse večje odpornosti bakterij proti protimikrobnim sredstvom zahteva, da imamo visoko tako specifičnost kot tudi senzitivnost. Za izboljšanje modela bi bilo potrebno povečati število izvidov in po možnosti razširiti obseg podatkov z rezultati laboratorijskih preiskav.

Ključne besede

Obdelava naravnega jezika, tekstovno rudarjenje, rezistentne bakterije, klasifikacija, strojno učenje, ESBL, Escherichia coli, zdravniški izvidi

Abstract

Title: Extraction of medical knowledge from a full-text description for predicting resistant bacteria infection

Resistant microorganisms are causing more and more problems in healthcare. Number of antibiotic-resistant microorganisms is growing faster than the number of newly discovered antibiotics. Wrongfully chosen antibiotics during treatment can also result in a greater resilience of the microorganisms. There exist several resistant microorganisms but we will focus on one, *Escherichia Coli* which produces ESBL enzymes. Patients usually start receiving proper antibiotic treatment when doctors get microbiological reports (which takes around two days). We try to predict if a patient has the previously mentioned bacteria *E. Coli* which produces ESBL enzymes, by using a medical report written in a natural language (which consists of the patient's history and status). Even if we achieved high specificity (90% with SVM and 86% with Naive Bayesian classifier) we can not use our models due to too low sensitivity (28% with SVM and 33% with Naive Bayesian classifier). Due to seriousness of the problem with resistant microorganisms it is required to have both metrics (specificity and sensitivity) high. In order to build better models we have to increase number of medical examination reports and maybe include additional results from other medical examinations.

Keywords

Natural language processing, text-mining, resistant bacteria, classification, machine learning, ESBL, Escherichia coli, medical examination report

Poglavje 1

Uvod

Odpornost bakterij proti protimikrobnim sredstvom vzbuja vse večjo skrb, saj se novi mehanizmi odpornosti pojavljajo in širijo po vsem svetu [1]. Število mikroorganizmov, odpornih proti protimikrobnim sredstvom, narašča hitreje od števila novih protimikrobnih sredstev. Nazadnje je bil nov razred protimikrobnih sredstev odkrit leta 1987 (slika 2.1). Posledično je s tem ogrožena sposobnost za zdravljenje pogostih in nalezljivih bolezni. Brez protimikrobnih sredstev za preprečevanje in zdravljenje okužb bi medicinski postopki, kot so presaditve organov, kemoterapije, vodenje sladkornih bolezni itd., postali zelo tvegani. Odpornost proti protimikrobnim sredstvom povečuje stroške zdravljenja, podaljšuje bivanje v bolnišnicah in potrebe po intenzivni negi. Odpornost proti protimikrobnim sredstvom tako rekoč izničuje stoletje razvoja (slika časovnice 2.1) protimikrobnih sredstev [31]. Poznamo tri osnovne mehanizme odpornosti proti protimikrobnim sredstvom [2]:

- encimska razgradnja protimikrobnih zdravil,
- sprememba bakterijskih proteinov, ki so tarča protimikrobnih sredstev,
- sprememba prepustnosti membrane za protimikrobna sredstva.

Osredotočili se bomo zgolj na encimsko razgradnjo v primeru bakterije *E. coli*, pozitivne na ESBL, ki jo najdemo v urinskem traktu.

V delu se bomo osredotočili na napovedovanje protimikrobne odpornosti pri posameznem bolniku. Vir podatkov za napovedovanje so medicinski izvidi, napisani v naravnem jeziku (slovenščini). Z obdelavo naravnega jezika, podatkovnega in tekstovnega rudarjenja bomo na podlagi anamneze in statusa bolnika poskušali napovedati prisotnost bakterije (odporne proti protimikrobnim sredstvom) pri bolniku. S tem bi lahko zmanjšali stopnjo tveganja pri predpisovanju protimikrobnih sredstev bolniku. Anamnezo in status

bolnika bi analizirali že ob sprejemu in tako ne bi bilo potreba dva dni čakati na izvide laboratorijske preiskave.

Z obdelavo naravnega jezika in s strojnim učenjem bomo določili uteži posameznih besed in poiskali povezave med pogostostjo določenih besed, njihovimi utežmi in s prisotnostjo bakterij, odpornih proti protimikrobnim sredstvom [3]. Besedilo izvida bo treba predhodno obdelati z metodami, kot so razčlenjevanje (tokenizacija), korenjenje besed, označevanje besed (samostalnik, veznik, ločilo itd.) in po potrebi brisanje določenih členov (npr. ločil). Preizkušali bomo več klasifikatorjev. Pomagali si bomo z orodjem Orange [32] in z lastno različico naivnega Bayesovega klasifikatorja, prilagojeno danemu problemu. Za svojo različico naivnega Bayesovega klasifikatorja smo se odločili zato, da smo lahko preizkušali različne funkcije za izračun pomembnosti besed. Dodali smo tudi možnost, da glede na priporočila iz literature izberemo ključne besede in jim določimo težo oz. množilnik vrednosti besede. Rezultate bomo primerjali s simulacijo odločanja zdravnika. Cilj je narediti model, ki bi znal napovedati boljše od simulacije zdravnikovega odločanja. Simulacijo zdravnikovega odločanja predstavlja naivni klasifikator, ki se odloča zgolj na podlagi pojavitve besede "ESBL" v izvidu. Pravilna klasifikacija je izvedena z mikrobiološkimi raziskavami, bolj podrobno opisano v poglavju 2.1.1.

V drugem poglavju opišemo pregled področja. Na kratko opišemo, kaj so encimi ESBL in kako delujejo, ter naštejemo dejavnike tveganja, ki povečajo verjetnost okužbe z bakterijo *E. coli*, ki je pozitivna na encime ESBL. Opišemo tudi laboratorijsko preiskavo, s katero se z gotovostjo lahko potrdi, ali je bakterija pozitivna na encime ESBL ali ne. Na kratko opišemo še bakterijo *E. coli* ter področje računalništva in matematike, ki se ukvarja z obdelavo naravnega jezika.

V tretjem poglavju opišemo metodologijo in standard CRISP-DM [20], po katerem je tudi potekal sam postopek našega dela. Po korakih opišemo postopek zbiranja podatkov (medicinskih izvidov).

V četrtem poglavju opišemo metode, ki smo jih uporabili. Opišemo metode in orodja, ki smo jih uporabili za obdelavo naravnega jezika, ter metode za klasifikacijo besedil. Podrobneje opišemo tudi vse metode za merjenje uspešnosti, ki smo jih uporabili.

V petem poglavju predstavimo rezultate za vsako metodo za klasifikacijo besedil posebej in nato še v tabeli predstavimo vse rezultate skupaj, ter podamo kvantitativno in kvalitativno analizo rezultatov.

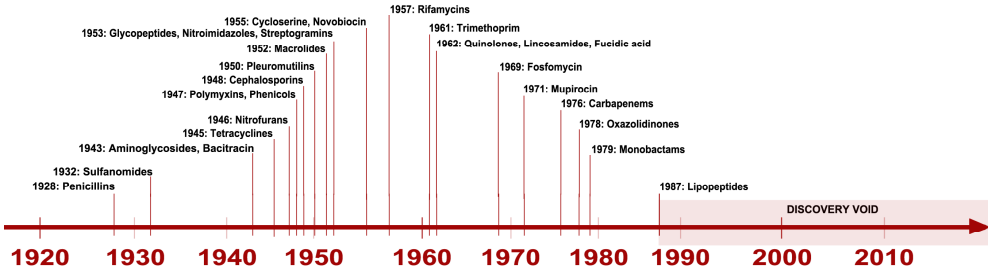
Poglavje 2

Pregled področja

2.1 Betalaktamaze razširjenega spektra

ESBL (angl. extended-spectrum beta-lactamases, slov. betalaktamaze razširjenega spektra) so encimi, ki povzročajo odpornost proti večini betalaktamskih protimikrobnih sredstev (npr. penicilini in cefalosporini) [5]. Encim razgradi betalaktamski obroč in s tem deaktivira protimikrobno sredstvo. Je glavni mehanizem obrambe pri Gram-negativnih bakterijah, odkrili so ga tudi pri nekaterih Gram-pozitivnih bakterijah. Če nastane okužba z bakterijo, ki proizvaja ESBL, bodo betalaktamska protimikrobna sredstva proti bakteriji neučinkovita, zato je okužbo potrebno zdraviti z drugimi protimikrobnimi sredstvi. Preventivno predpisovanju protimikrobnih sredstev, ki delujejo na ESBL, ni zaželeno, saj bi s tem protimikrobna sredstva po nepotrebnem izpostavljali bakterijam. V takšnih primerih bi omogočili hitrejši razvoj odpornosti proti protimikrobnim sredstvom. Najbolj pogosti bakteriji, ki proizvajata ESBL, sta Escherichia coli in Klebsiella pneumoniae [4]. Med dejavnike tveganja, ki povečajo verjetnost okužbe z bakterijo *E. Coli*, pozitivno na ESBL encime, spadajo [5][7]:

- starost nad 60 let,
- hospitalizacija,
- predhodna antibiotična terapija,
- pridružene bolezni,
- potovanje v tujino,
- prisotnost urinskega katetra,
- prisotnost nazogastrične sonde,
- prisotnost centralnega venskega katetra,
- nepokretnost,
- antibiotično zdravljenje kronične rane.



Slika 2.1: Odkritja novih razredov protimikrobnih sredstev skozi čas [8].

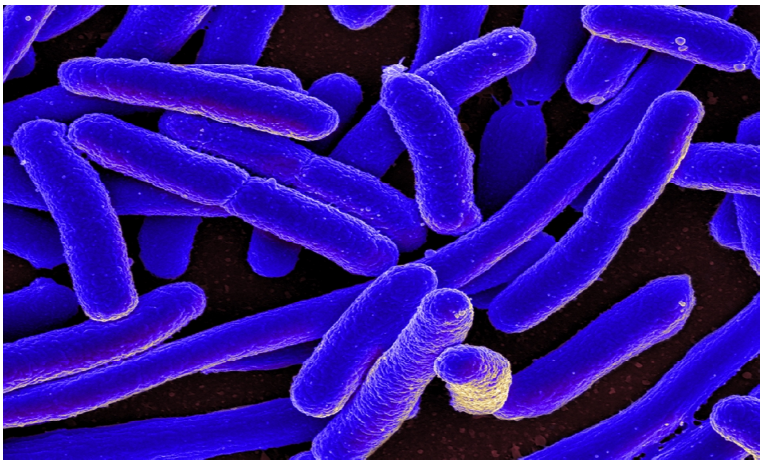
2.1.1 Testiranje odpornosti proti betalaktamom

V prvem koraku se izolirajo bakterije. Izolirane bakterije, v našem primeru *E. coli*, se nanese na agar (okolje, v katerem se bo bakterija razmnoževala), kjer bo izveden antibiogram. Antibiogram je test, kjer se na različne točke na agarju nanesejo različna protimikrobna sredstva in se opazuje, ali le ta delujejo na bakterijo. Če protimikrobna sredstva učinkujejo, nastane okoli tega sredstva kolobar. V primeru, ko je bakterija odporna proti protimikrobnim sredstvom razreda betalaktamov, je za potrditev odpornosti treba izvesti še sinergijski test z dvema diskoma (angl. double disc synergy test, DDST). Test deluje tako, da na agar nanesemo kužnino (vzorec za mikrobiološko preiskavo). Na sredino postavimo augmentin (mešanico med amoksicilinom in klavulansko kislino). Na eno stran poleg augmentina postavimo protimikrobno sredstvo ceftazidime, na drugo stran pa postavimo cefotaxime. Obe protimikrobni sredstvi pripadata razredu betalaktamov. Klavulanska kislina uniči encim betalaktamazo in tako lahko protimikrobni sredstvi začneta delovati. Če je okoli protimikrobnega sredstva nastal kolobar, pomeni, da je bakterija odporna proti protimikrobnim sredstvom iz razreda batalaktamov [6].

Slika 2.2: Agar z augmentinom na sredini in s protimikrobnima sredstvoma nad in pod njim. Pri zgornjem se lepo vidi nastali kolobar [6].

2.1.2 Escherichia coli

Escherichia coli (bolj poznana kot *E. coli*), je Gram-negativna enterobakterija rodu *Escherichis*. Prvi jo je leta 1885 opisal Theodor Escherich [9]. Je v večini primerov neškodljiva bakterija, ki prebiva v črevesju, lahko pa tudi povzroča okužbe in zastrupitev s hrano. Okužbe, ki jih povzroča *E. coli*, se navadno zdravijo s protimikrobnimi sredstvi, kot sta penicilin in cefalosporin. Toda v primerih, ko bakterija izloča encime ESBL, lahko povzroči okužbo, ki je ni mogoče pozdraviti s temi protimikrobnimi sredstvi. V takih primerih se zdravnik odloči za uporabo protimikrobnih sredstev, ki ne vsebujejo betalaktamskega obroča, saj so ti v tem primeru neučinkoviti [10] [30].



Slika 2.3: *Escherichia coli* pod elektronskim mikroskopom [29].

2.2 Obdelava naravnega jezika

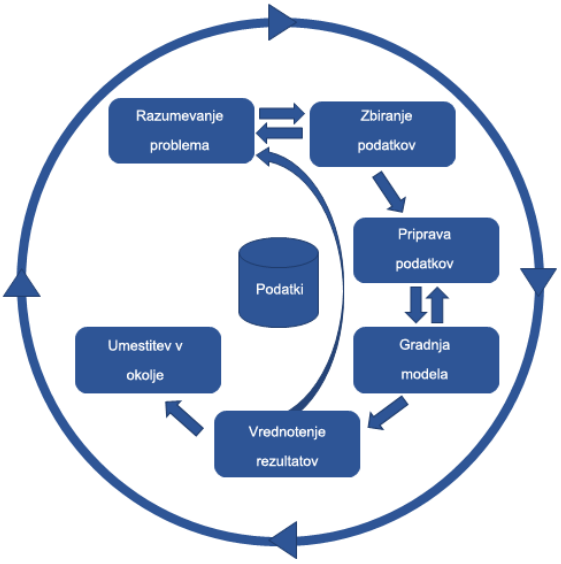
NLP (angl. Natural Language Processing, slov. obdelava naravnega jezika) je področje računalništva in matematike, ki raziskuje, kako uporabiti računalnike za razumevanje in manipuliranje naravnega jezika v tekstovni ali v govorni obliki. Vloga NLP v telekomunikacijah in računalništvu ni zanemarljiva. Tu so vključeni tudi sistemi za prepoznavo govora in za razumevanje jezika. Lahko se uporablja za izločanje podatkov, klasifikacijo, strojno prevajanje, izdelavo povzetkov in obdelavo besedila naravnega jezika. Primer izločanja so ekspertni sistemi za načrtovanje, ki izločijo pomembne podatke in jih shranijo za nadaljnjo uporabo [11]. Primer takšnega sistema je sistem, ki iz sporočila razbere, kdaj je sestanek, in ta podatek shrani v koledar, razbere telefonsko številko ipd. Primer uporabe obdelave naravnega jezika za pomoč pri klasifikaciji je tudi to magistrsko delo.

Poglavje 3

Metodologija

3.1 CRISP-DM

Sam potek dela je potekal po standardu CRISP-DM (slika 3.1) [20].



Slika 3.1: Predstavitev faz v modelu CRISP-DM.

V prvi fazi – razumevanje problema (angl. business understanding) smo preučili, kakšen problem imamo, kako ga bomo poskusili rešiti in na kakšne težave lahko naletimo. Problem, ki ga poskušamo rešiti, je, kako bolniku že na začetku hospitalizacije predpisati pravilno protimikrobno sredstvo (protimikrobno sredstvo, proti kateremu bakterija ni odporna). Če je iz izvida razvidno, da je bolnik okužen z bakterijo, pozitivno na ESBL, se mu lahko že takoj predpiše pravilno protimikrobno sredstvo. V nasprotnem primeru je treba počakati dva dni, da prispejo laboratorijski izvidi. Problem poskušamo rešiti tako, da s pomočjo strojnega učenja in tekstovnega rudarjenja analiziramo izvid ob sprejemu bolnika in poskušamo napovedati, ali je bolnik okužen z bakterijo, pozitivno na ESBL, ali ne. V tej fazi smo se tudi odločili, katera orodja bomo uporabili.

V drugi fazi – zbiranje podatkov (angl. data understanding) smo zbirali podatke (medicinske izvide). Za naš primer sta ključna dva dela izvida, in sicer anamneza in status. Zbiranje izvidov je potekalo na Kliniki za infekcijske bolezni in vročinska stanja. Med zbiranjem izvidov smo naleteli na več težav. Pri nekaterih bolnikih smo opazili, da njihova kri in urin nista bila oddana v laboratorij isti dan, kot so bili hospitalizirani. Take primere smo preskočili, saj se ni dalo zagotoviti, da so v času sprejema bili že okuženi. Opazili smo tudi, da so nekateri izvidi zelo kratki, in nas je skrbelo, ali mogoče vsebujejo premalo podatkov za učenje klasifikatorja in za samo klasifikacijo, vendar smo jih na koncu vseeno dodali v našo množico podatkov. Pri samem zbiranju izvidov smo tudi pazili na razmerje med številom bolnikov, okuženih z bakterijo pozitivno na ESBL in med številom bolnikov, ki niso okuženi. Poskušali smo preslikati razmerje iz populacije v Sloveniji na našo množico podatkov, to je 10% okuženih bolnikov in 90% bolnikov, ki niso okuženi. Zaradi nizkega odstotka okuženih bolnikov smo zbrali še nekaj izvidov za rezervo. Če bi imeli težave (bodisi tehnične ali človeška napaka) pri kakšnem izvidu, bi ga lahko zamenjali.

V tretjem koraku – priprava podatkov (angl. data preparation), je bilo besedila izvidov treba preoblikovati. Za obdelavo besedila smo uporabili orodja NLTK [34], Snowball [35] in ReLDI [36]. Orodja je bilo treba predhodno prilagoditi slovenščini. Označevalnik ReLDI smo nastavili s pomočjo predpripravljenih konfiguracij in slovarjev [37]. S pomočjo orodja ReLDI smo označili besede z značkami (veznik, ločilo...). Besede smo korenili s pomočjo orodij NLTK in Snowball. Za razčlenitev besedila na člene oz. tokene smo uporabili NLTK-jev Plaintext Corpus Reader [38]. Različico metode uteženega multinomialnega naivnega Bayesovega klasifikatorja, ki je bolj prilagojena za naš problem, smo razvili sami. Za druge metode smo uporabili orodje Orange [32]. Funkcijo za preoblikovanje besedila v matriko, ki predstavlja število pojavitev posameznih besed, smo razvili sami [41]. Podrobneje je priprava podatkov opisana v poglavju 4.4.1.

Sledi korak – gradnja modela (angl. modeling). Tu smo že začeli s strojnim učenjem. Preizkušali smo različne klasifikatorje z različnimi konfiguracijami. Občasno smo se tudi vračali na tretji korak (priprava podatkov), kjer smo znova pripravili podatke v drugačni

obliki (npr. v obliki matrike za potrebe orodja Orange).

V zadnjem koraku – vrednotenje rezultatov (angl. evaluation) smo primerjali rezultate napovedi klasifikatorjev. Na tej točki smo ovrednotili najboljše modele in prišli do zaključka, da je za uporabo v praksi treba izboljšati model.

3.2 Zbiranje podatkov

Zbiranje podatkov je potekalo na Kliniki za infekcijske bolezni in vročinska stanja v Ljubljani. Ker so podatki občutljive narave, smo morali pridobiti dovoljenje s strani Komisije Republike Slovenije za medicinsko etiko. Zbiranje je potekalo pod nadzorom prof. dr. Bojane Beovič, dr. med.

Predstavitev korakov zbiranja podatkov.

1. Postopek zbiranja podatkov se je začel tako, da smo na Inštitut za mikrobiologijo in imunologijo poslali prošnjo za seznam bolnikov. Seznam, ki smo ga prejeli, je bil v obliki tabele, v kateri so bili podatki o bolniku (slika 3.2).

2. Ročno smo pregeldali seznam, vnesli ime in priimek bolnika v program in pogledali, ali se datum sprejema bolnika ujema z datumom odvzema krvi in urina.

3. Če se je datum ujemal, smo imeli zagotovilo, da če je bolnik okužen z bakterijo *E. coli*, pozitivno na ESBL, jo je imel že pred prihodom na kliniko. Vse primere, pri katerih se datumi niso ujemali, smo preskočili.

4. Podatke o bolniku smo nato anonimizirane shranili v podatkovno bazo, pozneje namenjeno podatkovnemu rudarjenju.

Pri zbiranju izvidov smo pazili, da smo upoštevali razmerje iz populacije (10% bolnikov je okuženih z bakterijo, ki je pozitivna na ESBL). Ker je bilo delo zamudno in počasno, smo zaradi omejitve s časom, ki ga imamo na voljo, zbrali zgolj 210 izvidov. Od teh 210 jih je 11 (10+1, da ohrani razmerji 10% in 90%) dodatnih, namenjenih za rezervo, če nastane kakšna človeška ali tehnična napaka. Za raziskovanje in testiranje smo jih uporabili 199 (178 za neokužene bolnike in 21 za okužene bolnike).

Ime	Priimek	Datum rojstva	Kraj rojstva	Datum odvzema krvi in urina	ESBL
Janez	Novak	1.1.1970	Ljubljana	1.1.1990	Da
Ana	Novak	1.1.1970	Ljubljana	1.1.1990	Ne

Slika 3.2: Seznam bolnikov z njihovimi podatki. Podatki na tej sliki so zaradi varovanja osebnih podatkov izmišljeni.

```
Družinska anamneza: ni podatka.
Otroške bolezni: ni podatka.
Ostale dosedanje bolezni: gospod se zdravi zaradi karcinoma
prostate, osteoporoza, glavkoma. Ima obsežen ulkus leve
goleni. Stanje po zlomu desne stegnenice, stanje po
poškodbi hrbtenice (po poškodbi je še lahko hodil).
Alergični pojavi: nima.
Sedanja bolezen: danes zjutraj so mu v
DSO namerili vročino 39,2 st.C,
hropel je, saturacija 88%, RR 120/80, pulz 140/min.
Zdravnica mu je poslušala pljuča,
ugotavljala je inspiratorne pokce levo spodaj.
Že zjutraj je dobil 1. odmerek moksifloksacina 400 mg ter
Analgin subkutano ter kisik preko nosnega katetra.
Opoldne se mu je stanje poslabšalo, postal je
hipotenziven (RR 75/50), slabše odziven, padla je saturacija,
krvni sladkor ob tem 6.3.
Klicali so rešilca in gospoda pripeljali k nam.
V reševalnem vozilu je dobil 1l FR,
po čemer je tlak narastel na 96/56,
saturacija na 4l kisika preko nosnega katetra je bila 95%.
Gospod je bil sicer že pred 1 mesecem hospitaliziran
pri nas zaradi sepse, prejemal je najprej Ciprofloksazin,
nato Tazocin in nazadnje Linezolid.
V izboljšanem stanju so ga 18.5. odpustili nazaj v DSO.
Takrat naj bi bil izvor okužbe ulkus na goleni.
Od 21. - 30.11. je prejemal Ciprobay
zaradi suma na uroinfekt.
Epidemiološka anamneza: ni znana.
Socialna anamneza: oskrbovanec DSO.
Redna zdravila: Aspirin P 1 tbl zjutraj,
Bicalutamide teva 150 mg 1 tbl zjutraj,
Kalcijev karbonat 1g 1 tbl zjutraj,
Lexaurin 1 tbl po potrebi zvečer,
Mirzaten 30 mg 1 tbl zvečer.
Omnic ocas 4 mg 1 tbl zjutraj,
Plivit D3 35 kapljic vsako jutro,
Xalatan kapljice za oko 1x1 kapljica,
Acipan 20 mg zjutraj, Folacin 5 mg 1 tbl zjutraj,
Paladon 8 mg 1 cps zjutraj + 2 cps zvečer,
Metamizol stada (peroralne kapljice) 3x40 kapljic dnevno,
Lekadol 2 tbl zjutraj.
Vsadki, katetri: ni znano.
Kolonizacija z VOB: kolonizacija z MRSA,
izolirana iz nadzornih brisov v začetku novembra pri nas.
Izvide, ki jih prinaša s seboj: lab. izvidi 12.12.2017:
Hb 105, L 9.9 (seg. 89.9), CRP 60.

Klinični status ob pregledu:
Ob pregledu je gospod pri zavesti, delno orientiran,
```

```
sodelujoč, blago tahidispnoičen v mirovanju, afebrilen,
acianotičen, anikteričen, RR: 86/56, pulz: 82/min,
temp.: 36,1 st.C, saturacija: 90%.
Koža topla, suha. Bezgavke niso tipno povečane.
Glava normocefalna, žrelo bp, nos prehoden.
Vrat mehak, neboleč. Prsni koš simetričen, pomičen.
Nad pljuči so difuzno slišni hropci,
levo bazalno inspiratorni pokci.
Srčna akcija ritmična, tona tišja, šumov ne slišimm.
Trebuh v nivoju prsnega koša, mehak, palpatorno neboleč,
brez tipnih povečanih organov, vidna je brazgotina po
operaciji (odvzem kože za rekonstrukcijo palca desne roke),
vidna tudi velika, neukleščena,
neboleča ingvinalna hernija desno. Levo ingvinalno
tipna valjasta, premakljiva rezistenca velikosti 2x1 cm,
neboleča, koža nad njo bp. Ledveni poklep neboleč.
Okončine brez edemov, na levi goleni viden obsežen ulkus,
z rumenkasto-zelenkastimi oblogami, radialna in femoralna
pulza simetrično tipna. Meningealni znaki negativni.
ŽNS: zenici ožji, na levem očesu prisotna precejšnja katarakta,
desna zenica bp, bulbomotorika bp.
Z vsemi 4 okončinami normalno giba,
test na latentno parezo negativen.
```

Primer izvida, kjer je bil bolnik pozitiven na ESBL. Prisotnost ESBL pozitivne bakterije je bila potrjena na Inštitutu za mikrobiologijo in imunologijo s pomočjo laboratorijskih preiskav (dva dni po sprejemu bolnika).

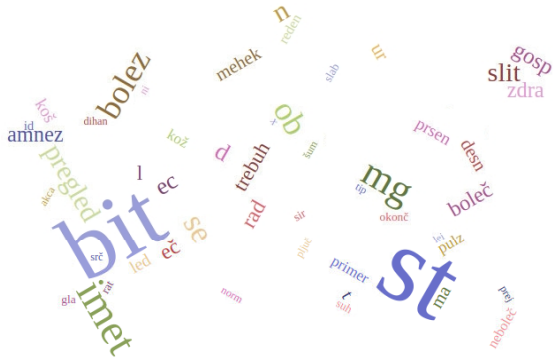
3.2.1 Analiza izvidov

Po uspešno zbranih izvidih (na izvornem oz. neobdelanem besedilu), smo opravili manjšo analizo besed, in sicer relativno frekvenco besed znotraj posameznega razreda (sliki 3.3 in 3.4). Analizirali smo tudi, kako se veča velikost zbirke obdelanih besed (več o obdelavi besedila v poglavju 4.2) z večanjem števila izvidov (sliki 3.5 in 3.6). Z izbiranjem značilk s pomočjo metode ANOVA (poglavje 4.6.9) smo poiskali, katere obdelane besede so najpomembnejše oz. imajo največjo težo (tabela 3.2).
Ker smo za nekatere obdelane besede dobili "enočrkovne besede", smo si pogledali kakšne so bile na začetku (tabela 3.1).

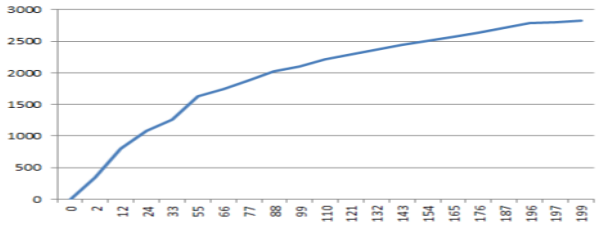
st:	4.47%	j:	0.73%	mehek:	0.51%	stan:	0.38%
bit:	4.12%	gosp:	0.68%	pulz:	0.51%	primer:	0.38%
mg:	1.57%	l:	0.68%	zjutraj:	0.51%	slab:	0.38%
imet:	1.13%	d:	0.65%	trebuh:	0.49%	tipen:	0.37%
ob:	1.08%	zdra:	0.65%	led:	0.49%	simetrič:	0.37%
pregled:	0.98%	t:	0.59%	neboleč:	0.47%	zer:	0.37%
on:	0.96%	amnez:	0.58%	dihan:	0.47%	akca:	0.37%
n:	0.91%	ma:	0.56%	sir:	0.47%	pljuč:	0.37%
bolez:	0.91%	x:	0.56%	ec:	0.45%	seč:	0.37%
ur:	0.85%	tbet:	0.56%	prsen:	0.45%	šum:	0.37%
se:	0.82%	id:	0.56%	boleč:	0.45%	zji:	0.37%

Slika 3.3: Grafični prikaz 52 najbolj pogostih obdelanih besed v izvidih, kjer je bil bolnik okužen z bakterijo. Večja velikost pisave pomeni višjo relativno frekvenco.

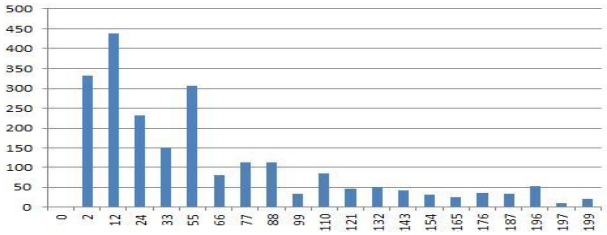
st:	4.76%	amnez:	0.65%	prsen:	0.52%	gla:	0.40%
bit:	3.92%	zdra:	0.65%	led:	0.52%	šum:	0.40%
imet:	1.33%	eč:	0.64%	pulz:	0.51%	dihan:	0.40%
mg:	1.20%	boleč:	0.64%	kož:	0.49%	akca:	0.40%
on:	1.11%	rad:	0.62%	primer:	0.49%	ni:	0.40%
ob:	0.98%	trebuh:	0.61%	neboleč:	0.46%	srč:	0.40%
bolez:	0.97%	l:	0.61%	reden:	0.45%	norm:	0.40%
se:	0.95%	ur:	0.58%	id:	0.44%	prej:	0.40%
pregled:	0.82%	ma:	0.57%	okonč:	0.42%	tip:	0.39%
n:	0.81%	mehek:	0.55%	slab:	0.41%	pljuč:	0.39%
slit:	0.79%	t:	0.55%	x:	0.41%	lei:	0.38%



Slika 3.4: Grafični prikaz 52 najbolj pogostih obdelanih besed v izvidih, kjer bolnik ni bil okužen z bakterijo. Večja velikost pisave pomeni višjo relativno frekvenco.



Slika 3.5: Graf prikazuje spreminjane velikosti zbirke obdelanih besed glede na število izvidov. Opazimo lahko, da raste vedno počasneje.



Slika 3.6: Graf prikazuje spreminjanje števila na novo dodanih obdelanih besed v zbirko, glede na število izvidov.

Obdelana beseda	**Seznam možnih izvornih besed**
d	d
l	l
x	x, xa
j	je

Tabela 3.1: Nekaj primerov besed, ki so po obdelavi nastale enočrkovne.

Beseda	**Ocena**	**Beseda**	**Ocena**	**Beseda**	**Ocena**
razlož	18.549	kolez	14.093	sanfor	10.874
fenobarbiton	18.549	ozek	14.093	delen	10.866
utrip	18.549	tak	13.937	besed	10.642
tra	18.549	oseb	12.623	sirlud	10.600
ol	16.669	obsež	12.522	gnoja	10.600
presadek	16.520	j	11.417	erb	10.600
odzet	14.612	jaz	11.105	odr	10.600

Tabela 3.2: Tabela prikazuje 21 obdelanih besed (več o obdelavi v poglavju 4.2) z največjo težo in njihovimi ocenami. Besede so bile pridobljene s pomočjo metode ANOVA.

Poglavje 4

Pregled metod

4.1 Obdelava naravnega jezika

Obdelava naravnega jezika je sestavljena iz več korakov: tokenizacija ali razčlenitev (razbitje besedila na tokene oz. člene), korenjenje besed (npr. besedo "pisava" preoblikuje v besedo "pis") in označevanje besed (npr. "avto" označi kot samostalnik in "," označi kot znak). Pred samim učenjem klasifikatorja in pozneje pred klasifikacijo besedila lahko besedilo še dodatno preoblikujemo. Po vseh prej omenjenih korakih lahko iz besedila pobrišemo še veznike in znake.

4.1.1 Označevanje besedila

Za označevanje besedila smo uporabili označevalnik besedila ReLDI [36]. Označevalnik besedilo razdeli na člene (besede ali znake) in jim dodeli značke. Nekaj primerov značk je v tabeli 4.1.

Značka	**Pomen**	**Značka**	**Pomen**
Z	ločilo	N	samostalnik
Cc	priredni veznik	V	glagol
Cs	podredni veznik	A	pridevnik

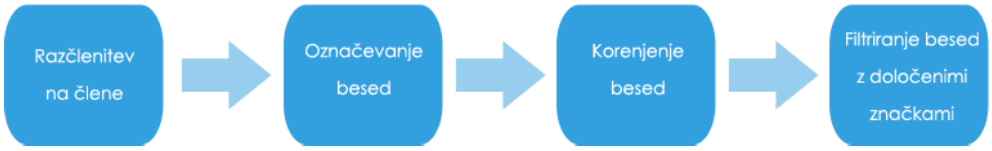
Tabela 4.1: Primeri značk s katerimi označevalnik ReLDI označi besede oz. znake [25].

4.2 Predhodna obdelava besedila

Besedilo smo pred učenjem klasifikatorja in pred klasifikacijo predhodno obdelali. V besedilu smo popravili tiskarske napake, ki smo jih odkrili med analiziranjem izvidov (tabela 4.1), in iz besedila izbrisali vse pojavitve besede "ESBL". Besedilo smo nato razčlenili na člene, jih označili, korenili, in nato pobrisali veznike in znake (na podlagi značk).

Beseda v izvidu	**Popravljena beseda**
Alzhaimer	Alzheimer
askultatur	avskultator
bronhopnemonija	bronhopnevmonija
ESLB	ESBL
tahikradičen	tahikardičen

Slika 4.1: Nekaj primerov tiskarskih napak in njihovih popravkov.



Slika 4.2: Predstavitev korakov obdelave besedila pred učenjem klasifikatorja in napovedovanjem.

Predstavitev korakov na izseku besedila iz izvida je na sliki 4.3.

Začetno oz. vhodno besedilo.
Ostale dosedanje bolezni: gospod se zdravi zaradi karcinoma prostate, osteoporoza, glavkoma.

Po razdelitvi besedila na člene oz. tokene.
['Ostale', 'dosedanje', 'bolezni', ':', 'gospod', 'se', 'zdravi', 'zaradi', 'karcinoma', 'prostate', ',', 'osteoporoza', ',', 'glavkoma', '.']

Rezultat po korenjenju.
['ostal', 'dosed', 'bolez', ':', 'gospod', 'se', 'zdrav', 'zarad', 'karc', 'prost', ',', 'osteoporoz', ',', 'glavk', '.']

Rezultat po filtriranju besed z določenimi značkami.
['ostal', 'dosed', 'bolez', 'gospod', 'zdrav', 'zarad', 'karc', 'prost', 'osteoporoz', 'glavk']

Slika 4.3: Predstavitev korakov obdelave besedila na izseku iz medicinskega izvida.

4.3 Testne metodologije

Za testiranje smo množico podatkov razdelili na tri dele, na učno, validacijsko in testno množico. To smo dosegli tako, da smo uporabili dvojno, ugnezdeno poravnano 10-kratno prečno preverjanje. Preizkusili smo tudi metodo "izpusti enega" (angl. leave-one-out), vendar je ta predolgo trajala za časovni okvir, ki ga imamo na voljo.

Opis postopka gradnje množic podatkov, izbiranja parametrov in hiperparametrov, validacije in testiranja:

1. Množico razdelimo na k disjunktnih podmnožic (v našem primeru k=10). Ker uporabljamo poravnano prečno preverjanje, je razmerje med vrednostimi razredov znotraj podmnožic enako. *k-1* podmnožic uporabimo za učenje, ena podmnožica je namenjena testiranju.

2. Ker je za izbiranje parametrov pomembno, da to delamo samo na učni množici, je treba te *k-1* podmnožice še naprej razdeliti. Razdelimo jih na enak način kot v koraku 1.

3. Naj bo $j = k - 1$, na *j-1* podmnožicah se učimo, na eni podmnožici pa preverimo izbiro parametrov. To ponovimo *j-krat*. V tem koraku izbiramo parametre, kot so uteži besed za uteženi multinomialni naivni Bayesov klasifikator, število korakov pri SGD klasifikatorju itd. Pri izbiri hiperparametra pa smo se osredotočili samo na število značilk. Značilke smo izbirali s pomočjo metode ANOVA in sicer 99 najpomembnejših značilk (oz. besed z največjo težo). Z izbranim parametrom in hiperparametrom, preverimo model na validacijski množici (ena podmnožica izmed j podmnožic). Na isti validacijski množici smo testirali več različnih parametrov in hiperparametrov. Najboljšo konfiguracijo oz. izbiro parametrov in hiperparametrov smo si shranili. Po vseh j iteracijah smo pogledali, katera izmed konfiguracij se največkrat pojavlja, in s takšno konfiguracijo smo nato testirali na testni množici (ena podmnožica pri *k-1* podmnožicah). Za boljšo konfiguracijo smo izbrali tisto, ki je imela večjo vsoto (senzitivnost + specifičnost).

4. Po zaključenih k iteracijah izračunamo povprečje rezultatov napovedi (mere uspešnosti so opisane v poglavju 4.6).

4.4 Metode iz orodja Orange

Orodje Orange [32] je bilo razvito na Univerzi v Ljubljani v okviru Laboratorija za bioinformatiko na Fakulteti za računalništvo in informatiko. Laboratorij se ukvarja s podatkovno analitiko, strojnim učenjem in zlivanjem podatkov. Orodje Orange je namenjeno analizi in vizualizaciji podatkov.

4.4.1 Priprava podatkov

Orodje Orange sprejema vhodne podatke v obliki matrike, zato je bilo besedilo iz naravnega jezika v tekstovni obliki treba pretvoriti v matriko. Besedilo izvidov smo predhodno preoblikovali po enakem postopku kot v razdelku 4.2. Iz tako preoblikovanega besedila smo zgradili unijo obdelanih besed in prešteli njihove pojavitve v vsakem besedilu (izvidu). Stolpci so predstavljali število pojavitev (frekvenco) posamezne obdelane besede v izvidu, vrstice so predstavljale izvide. Razred, ki smo ga napovedovali, je bil *ESBL*, ki je imel diskretno vrednost 0 (laboratorijski testi niso pokazali prisotnosti ESBL) ali 1 (laboratorijski testi so pokazali prisotnost ESBL). Tako pripravljeno matriko smo uporabili v orodju Orange [32]. Če je kateri izmed stolpcev imel konstantno vrednost, smo ga odstranili.

Primer preoblikovanja besedila v vektor je predstavljen na sliki 4.4. Matrika, ki smo jo na koncu dobili, je imela 200 vrstic (kar predstavlja 200 primerov) in približno 2900 stolpcev (kar predstavlja 2900 različnih besed). Če besedila ne bi preoblikovali (npr. besedi "poškodba" in "poškodovan" se po preoblikovanju spremenita v "poškod"), bi bile dimenzije matrike še večje (6800 stolpcev).

Izvirno besedilo:

Ima obsežen ulkus leve goleni. Stanje po zlomu desne stegnenice, stanje po poškodbi hrbtenice (po poškodbi je še lahko hodil).

Vektor:

po	stegnenice	še	poškodbi	hodil	desne	lahko	stanje	ima	goleni	ulkus	obsežen	hrbtenice	je	zlomu	leve
3	1	1	2	1	1	1	2	1	1	1	1	1	1	1	1

Slika 4.4: Vektor, kjer številka predstavlja število pojavitev posamezne besede v povedi.

4.4.2 Metode

V orodju Orange smo uporabili metode SVM [12], nevronsko mrežo [13], naivni Bayesov klasifikator [15], logistično regresijo [14], SGD [16] in klasifikacijska pravila CN2 [17].

4.4.3 SVM

Pri metodi podpornih vektorjev (SVM) smo poskušali z različnimi jedrnimi funkcijami (linearno jedro, sigmoid in RBF) in z različnimi vrednostmi parametra c. Za vrednosti parametra c smo izbrali vrednosti 0.1, 1, 10, 100. Takšne vrednosti smo izbrali zato, ker smo dobivali slabe rezultate in smo hoteli poiskati razred velikosti parametra c, kjer bi bile opazne spremembe v napovedih. Za najboljšo se je izkazala izbira sigmoidne jedrne funkcije, za najslabšo pa izbira jedrne funkcije RBF.

Jedrne metode

Jedro je funkcija, ki preoblikuje prostor atributov v nov prostor (npr. iz nelinearnega prostora v linearen prostor) in s tem se lažje poišče ločnico med razredi [18].

Jedro RBF (z drugim imenom Gaussovo jedro), preslika n-dimenzionalen prostor v neskončno dimenzionalen prostor. Če imamo k točk, namenjenih učenju, nam to jedro omogoča, da SVM deluje v k dimenzionalnem prostoru.

Linearno jedro je najbolj preprosta jedrna funkcija. Uporaba linearne jedrne funkcije ohrani število dimenzij.

Uporaba sigmoidne jedrne funkcije skupaj z metodo podpornih vektorjev je enakovredno nevronski mreži z dvema nivojema (brez skritih nivojev) [24].

4.4.4 Logistična regresija

Pri logistični regresiji se je boljše izkazala uporaba regularizacije L2 (formula 4.2). Preizkušali smo tudi regularizacijo L1 (formula 4.1).

$$\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \Delta\sum_{j=i}^{n}|\theta_j| \tag{4.1}$$

Logistična regresija z regularizacijo L1 [26].

$$\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \Delta\sum_{j=i}^{n}\theta_j^2 \tag{4.2}$$

Logistična regresija z regularizacijo L2 [26].

Regularizacija

Regularizacija je zelo pomembna tehnika v strojnem učenju. Namenjena je preprečevanju prekomernega prilagajanja podatkom (angl. overfitting). Težava prekomernega prilagajanja je, da se model tesno prilagaja učni množici podatkov in ni dovolj robusten [27].

Regularizacija L1 (poimenovana tudi LASSO) prišteje kazen enako vsoti absolutnih vrednosti koeficientov (formula 4.3) [27].

$$Error_{L1} = Error \ + \Delta \sum_{i=1}^{N} |\theta_i| \tag{4.3}$$

Formula regularizacije L1, kjer θ_i predstavlja parameter in Δ določa težo regularizacije.

Regularizacija L2 (poimenovana tudi Ridge) prišteje kazen enako vsoti kvadratov koeficientov (formula 4.4) [27].

$$Error_{L2} = Error \ + \Delta \sum_{i=1}^{N} \theta_i^2 \tag{4.4}$$

Formula regularizacije L2, kjer θ_i predstavlja parameter in Δ določa težo regularizacije.

4.4.5 CN2

Pri klasifikacijskih pravilih CN2 smo uporabili dve ocenjevalni funkciji za iskanje pravil (Entropy in Laplace Accuracy). Za boljšo se je izkazala uporaba funkcije Laplace(uporaba Laplaceove metode).

4.4.6 SGD

Uporabili smo tudi klasifikator SGD. Najboljše rezultate smo dobili s kombinacijo optimizacijske metode statističnega gradientnega sestopa (angl. Stochastic gradient descent) z linearno funkcijo in regularizacijo elastic net.

Regularizacija elastic net

Regularizacija elastic net je regularizacijska metoda, ki linearno združuje regularizacijski metodi L1 (Lasso) in L2 (Ridge) (formula 4.5). Težo regularizacije določa parameter

Δ, težo posamezne regularizacijske metode pa določa parameter α ($\alpha \in (0,1)$). Če eni od regularizacijskih metod dodelimo večjo težjo, posledično s tem zmanjšamo težo drugi regularizacijski metodi.

- $\alpha > 0.5$, večjo težo ima regularizacija L1.
- $\alpha < 0.5$, večjo težo ima regularizacija L2.
- $\alpha = 0.5$, obe regularizaciji sta enakovredni.

$$Error_{enet} = Error \ + \Delta((1-\alpha)\sum_{i=1}^{N}\theta_i^2 \ + \ \alpha\sum_{i=1}^{N}|\theta_i|) \tag{4.5}$$

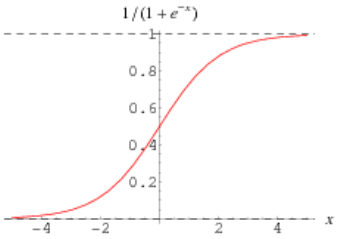
4.4.7 Nevronska mreža

Preizkušali smo različne globine, različna števila nevronov in različne aktivacijske funkcije. Najboljše rezultate smo dosegli z dvema skritima nivojema. Za aktivacijsko funkcijo smo uporabili sigmoidno funkcijo (ker smo želeli imeti omejitev izhodnih podatkov iz nevrona). Uteži smo računali s pomočjo metode L-BFGS-B.

Sigmoidna funkcija

Je ena izmed logističnih funkcij z značilno krivuljo v obliki črke "S" (slika 4.5). Namenjena je normaliziranju vhodne spremenljivke med 0 in 1 ali med -1 in 1. Formula sigmoidne funkcije je v enačbi 4.6.

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \tag{4.6}$$



Slika 4.5: Graf sigmoidne funkcije na intervalu (-5,5) [33].

Metoda BFGS

Metoda se uporablja za iskanje stacionarne točke oz. natančneje iskanje minimuma funkcije. Poimenovana je po matematikih Charlesu Georgeu Broydenu, Rogerju Fletcherju, Donaldu Goldfarbu in Davidu Shannoju (Broyden-Fletcher-Goldfarb-Shanno). Metoda spada med kvazi-Newtonove metode. Različica L-BFGS-B je namenjena primeru, ko želimo omejiti porabo pomnilnika (**L**: limited memory) in imamo določeno spodnjo in zgornjo mejo (l < x < u) (**B**: bound-constrained) [19].

4.5 Uteženi multinomialni naivni Bayesov klasifikator

Med zbiranjem podatkov (izvidov) smo opazili, da so nekateri izvidi izredno kratki (bodisi so jih napisali študentje bodisi bolnik ni navajal anamneze). Nekateri izvidi so bili od trido štirikrat krajši od najdaljših in so posledično vsebovali veliko manj podatkov. Rešujemo tudi zelo specifičen problem, za katerega vemo, da določeni kriteriji povečajo verjetnost, da je bolnik okužen z bakterijo, pozitivno na ESBL (dejavniki tveganja omenjeni v poglavju 2.1). Imamo lahko dve skrajnosti: zelo kratek izvid, ki vsebuje veliko pomembnih podatkov (npr. podatki o dejavnikih tveganja) ali primer dolgega izvida, ko večino izvida sestavlja anamneza z zelo malo pomembnimi podatki.

Da bi rešili ta primer, smo se odločili narediti svojo različico klasifikatorja, ki omogoča dodajanje besednih uteži. Utežitev deluje tako, da določene besede (*hospitalizacija*, *kateter*, *nazogastrična*, *tujina*, *vob*) označimo kot pomembne (v kodi jih poimenujemo kot *keywords*) in jim dodelimo utež. Za te besede smo se odločili na podlagi dejavnikov, ki povečajo verjetnost okužbe z bakterijo *E. coli*, ki je pozitivna na encime ESBL [5][7].

Besedilo je bilo treba klasificirati v eno izmed dveh vrednosti razreda. Na voljo imamo en razred, ki ima dve diskretni vrednosti, je ali ni pozitiven (ESBL enak 1 ali ESBL enak 0). V fazi učenja klasifikatorja smo izračunali apriorne verjetnosti razredov $P(c)$ (formula 4.7) in pogojne verjetnosti (verjetnosti pojavitve posamezne besede za posamezen razred) $P(t|c)$ (formula 4.8). Pri računanju pogojne verjetnosti smo upoštevali prej omenjene uteži besed (formula 4.9). V fazi klasifikacije izračunamo, s kakšnimi verjetnostimi izvid pripada vsakemu izmed razredov $P(c|d)$ (formula 4.10). Izberemo razred, ki je imel najvišjo izračunano verjetnost (formula 4.11). Najboljši rezultat smo dobili, tako da smo besedam *vob*, *nepokretnost* in *bakterij*, določili utež 2.

4.5.1 Multinomialni naivni Bayesov klasifikator

V modelu multinomialnega naivnega Bayesovega klasifikatorja se predvideva, da so dolžine posameznih dokumentov neodvisne od razreda. Naivno je tudi predvidevanje, da je verjetnost vsake besede v dokumentu neodvisna od konteksta in položaja besede v dokumentu [28].

Naj bo

$P(c_k)$	-	apriorna verjetnost razredov r_k
N_c	-	število izvidov v razredu c
N	-	število vseh izvidov
$P(t\|c)$	-	pogojna verjetnost, da beseda t pripada razredu c
T_{ct}	-	število pojavitev besede t v razredu c
$\|V\|$	-	število unikatnih besed v učni množici
$\sum_{t' \in V} T_{ct'}$	-	število vseh besed v razredu c
$\hat{P}(c\|d)$	-	verjetnost, da dokument d pripada razredu c
$P(t_k\|c)$	-	verjetnost, da se beseda t_k pojavlja v razredu c
w_t	-	utež besede

Opis delovanja po korakih.

1. Izračun razmerja razredov (oziroma apriorne verjetnosti razreda) v učni množici (formula 4.7). *Primer: učna množica ima 10 primerov, razred A se pojavi 4-krat. Razmerje je torej* $\frac{4}{10}$.

$$\hat{P}(c) = \frac{N_c}{N} \tag{4.7}$$

2. Izračun pogojnih verjetnosti (formula 4.8) in izračun pogojnih verjetnosti z upoštevanjem uteži besed (formula 4.9). Da bi se izognili ničlam, smo uporabili Laplaceovo glajenje (prišteje 1 v števcu in $|V|$ v imenovalcu).

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + |V|)} \tag{4.8}$$

$$\hat{P}(t|c) = \frac{\boldsymbol{w_t} * T_{ct} + 1}{(\sum_{t' \in V} \boldsymbol{w_t} * T_{ct'}) + |V|)} \tag{4.9}$$

3. Izračun verjetnosti, da dokument d pripada razredu c (formula 4.10).

$$\hat{P}(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \tag{4.10}$$

4. Izbran razred je tisti, ki ima najvišjo izračunano verjetnost po formuli 4.10 (formula 4.11).

$$\hat{c} = \underset{c}{\operatorname{argmin}}\, \hat{P}(c|d) \tag{4.11}$$

4.6 Mere uspešnosti

Za oceno kakovosti modela oz. klasifikatorja smo uporabili različne metrike. Uporabili smo senzitivnost, specifičnost, preciznost, oceno F1, klasifikacijsko točnost in AUC. Te metrike smo izračunali na podlagi množice podatkov, namenjeni končnemu testiranju (podrobneje v poglavju 4.3). Primerjava klasifikatorjev na podlagi klasifikacijske točnosti se je izkazala za neuporabno. Razlog za to je razmerje med vrednostmi razreda (10% primerov, kjer je ESBL enak 1, in 90%, kjer je ESBL enak 0). Na prvi pogled 90% deluje kot zelo dobra ocena, ampak to lahko pomeni, da nismo niti enega primera, kjer je ESBL, enak

1, napovedali pravilno. Mi smo se za primerjavo med klasifikatorji osredotočili predvsem na senzitivnost in specifičnost. Pri primerjanju dveh klasifikatorjev, smo se odločili, da je boljši tisti, ki ima višjo senzitivnost in specifičnost.

4.6.1 Prečno preverjanje

Prečno preverjanje je neizčrpna metoda preverjanja. *k-kratno* prečno preverjanje razdeli množico na k enako velikih delov. Za učenje se uporabi *k-1* delov, preostali del se uporabi za testiranje. Če je prečno preverjanje poravnano, pomeni, da vsak izmed k delov vsebuje enako razmerje obeh vrednosti razreda (v našem primeru 10% ESBL-pozitivno, 90% ESBL-negativno).

4.6.2 Senzitivnost

Senzitivnost (angl. recall) predstavlja razmerje med napovedanimi pozitivnimi elementi in vsemi pozitivnimi elementi. Za naš primer je izračun senzitivnosti predstavljen v enačbi 4.12.

$$senzitivnost = \frac{(število\ pravilnih\ napovedi,\ kjer\ je\ razred\ ESBL\ enak\ 1)}{(število\ vseh\ primerov,\ kjer\ je\ razred\ ESBL\ enak\ 1)} \tag{4.12}$$

4.6.3 Specifičnost

Specifičnost (angl. specificity) predstavlja razmerje med napovedanimi pomembnimi negativnimi in vsemi negativnimi elementi. Za naš primer je izračun specifičnosti predstavljen v enačbi 4.13.

$$specifičnost = \frac{(število\ pravilnih\ napovedi,\ kjer\ je\ razred\ ESBL\ enak\ 0)}{(število\ vseh\ primerov,\ kjer\ je\ razred\ ESBL\ enak\ 0)} \tag{4.13}$$

4.6.4 Preciznost

Preciznost (angl. precision) predstavlja razmerje med pravilno napovedanimi pozitivnimi elementi in vsemi elementi, ki so napovedani kot pozitivni. Za naš primer je izračun preciznosti predstavljen v enačbi 4.14.

$$\begin{aligned} tp &= \check{s}tevilo\ pravilnih\ napovedi,\ kjer\ je\ razred\ ESBL\ enak\ 1 \\ fp &= \check{s}tevilo\ nepravilnih\ napovedi,\ kjer\ je\ pravilen\ razred\ ESBL\ enak\ 0 \\ preciznost &= \frac{tp}{tp+fp} \end{aligned} \tag{4.14}$$

Izračun preciznosti v našem primeru.

4.6.5 Ocena F1

Ocena F1 (angl. F1-score) je harmonično povprečje senzitivnosti in preciznosti. Če je eden izmed njiju enak 0, je ocena F1 tudi enaka 0 (formula 4.15).

$$ocena\ F1 = 2 * \frac{senzitivnost * preciznost}{senzitivnost + preciznost} \tag{4.15}$$

4.6.6 Klasifikacijska točnost

Klasifikacijska točnost (angl. classification accuracy) je delež pravilno napovedanih napovedi. V našem primeru smo morali biti pazljivi, saj ima 90% primerov razred ESBL enak 0, torej če vedno napovemo, da bolnik ni okužen z bakterijo, ki je pozitivna na ESBL, dobimo klasifikacijsko točnost 90%. Napoved je slaba, čeprav je na prvi pogled videti dobra. Klasifikacijska točnost ima večjo težo v primerih, ko obe vrednosti razreda predstavljata približno enak (ali enak) delež (50% ima vrednost 1 in 50% ima vrednost 0).

4.6.7 Krivulja ROC

Krivulja ROC (angl. Receiver Operating Characterstic curve) je grafična predstavitev kakovosti binarnega klasifikatorja (klasifikator, ki napoveduje en razred z dvema vrednostima) in nam omogoča analizo razmerja med senzitivnostjo in specifičnostjo (slika 4.6). Na abscisni osi je prikazano relativno število lažno pozitivnih napovedi (angl. false-positive rates), kar je *1-specifičnost*. Na ordinatni osi pa je prikazano relativno število resnično pozitivnih napovedi (angl. true-positive rates), kar je *senzitivnost*. Krivulja ima dve trivialni točki: točka, kjer je senzitivnost 1 in specifičnost 0, ter točka, kjer je senzitivnost 0 in specifičnost 1. Bolj ko je krivulja blizu diagonale, slabši je klasifikator in nasprotno, dlje kot je od diagonale, boljši je. Če je krivulja blizu spodnjega desnega kota (senzitivnost je

enaka 0, 1-specifičnost je enaka 1), lahko samo napovedi v binarnem klasifikatorju zamenjamo (npr. kar je bilo prej 1, je sedaj 0 in nasprotno). S tem krivuljo premaknemo nad diagonalo.

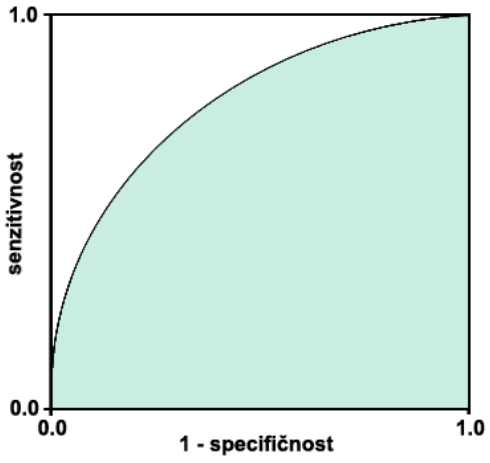
Vsak klasifikator je prikazan s točko, ki ustreza njegovi senzitivnosti in specifičnosti. Za vsak prag, ki ga izberemo za odločitveno pravilo, dobimo drugo točko na grafu. Običajno izberemo prag 0.5, kar pomeni, da klasificiramo primer v tisti razred, ki ima večjo verjetnost (verjetnost, večjo od 0.5). S spreminjanjem praga se premikamo po krivulji. Praga 0.0 in 1.0 predstavljata trivialni točki. Prag 0.0 pomeni, da vse primere klasificiramo kot negativne, 1.0 pa pomeni, da klasificiramo vse primere kot pozitivne. Diagonala, ki povezuje trivialni točki, ustreza naključnim klasifikatorjem [21].

4.6.8 AUC

AUC (angl. area under curve) je ploščina pod krivuljo ROC (slika 4.6). Iz ocene AUC lahko izpeljemo formulo za kakovost napovedi (formula 4.16).

$$kakovost = 2 * |0.5 - AUC| \tag{4.16}$$

Izračun kakovosti klasifikatorja. Rezultat je normaliziran med 0 in 1, kjer večje število predstavlja bolj kakovosten klasifikator.



Slika 4.6: Primer krivulje ROC in AUC (obarvana ploščina pod krivuljo).

4.6.9 ANOVA

Pri analiziranju besedila smo tudi poiskali besede z največjo težo. Pomagali smo si z metodo ocenjevanja atributov, ANOVA (analiza variance, angl. **an**alysis **o**f **va**riance). Za ocenjevanje smo uporabili matriko s frekvencami besed (več o matriki v poglavju 4.4.1). Izračunali smo oceno F (angl. F-value) med vsakim atributom in razredom (ESBL). Višja ocena atributa pomeni, da atribut boljše opisuje razred. Da dobimo 20 najboljših atributov, vzamemo 20 atributov z najvišjo oceno F [22][23].

Naj bo

- Y - skupno povprečje
- K - število skupin
- N - skupna velikost vzorca
- Y_i - povprečje v *i-ti* skupini
- n_i - število opazovanj v *i-ti* skupini

1. Najprej izračunamo varianco med skupinami (npr. med atributom in razredom), kar je prikazano v formuli 4.17.

$$\sum_{i=1}^{K} n_i(\overline{Y}_i - \overline{Y})^2/(K-1) \tag{4.17}$$

2. V drugem koraku izračunamo varianco med elementi znotraj skupine (formula 4.18).

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2/(N-K) \tag{4.18}$$

3. Oceno F dobimo tako, da varianco med skupinami delimo z varianco med elementi znotraj skupine (formula 4.19).

$$F = \frac{\sum_{i=1}^{K} n_i(\overline{Y}_i - \overline{Y})^2/(K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2/(N-K)} \tag{4.19}$$

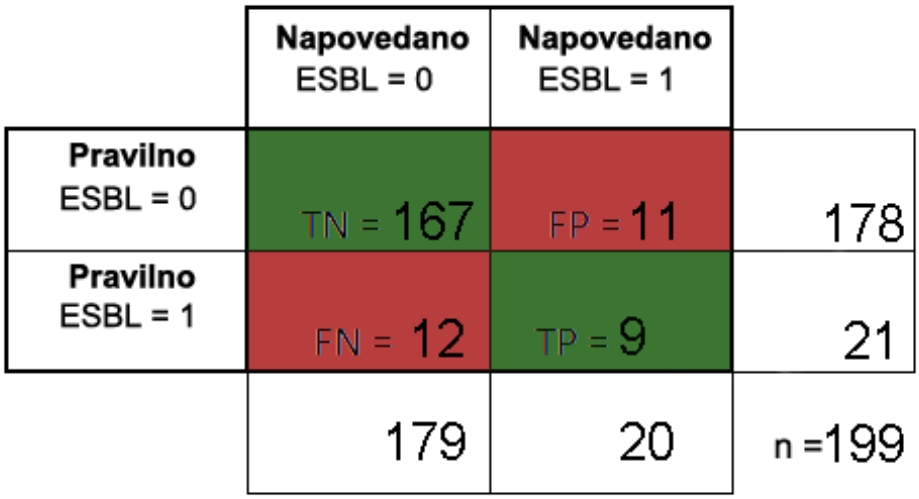
Poglavje 5

Rezultati

Za lažjo primerjavo smo v sodelovanju z zaposlenimi na Kliniki za infekcijske bolezni in vročinska stanja, pripravili model, ki simulira odločanje zdravnika. Naše metode smo primerjali s tem modelom. Model, ki smo ga naredili, je sicer preprost. Klasifikator napove "ESBL" samo v primeru, ko se v besedilu pojavi beseda "ESBL". Posledično smo dali večjo težo specifičnosti. Nas je v osnovi zanimalo, ali lahko napovemo, da ima bolnik bakterijo *E. coli*, ki je pozitivna na ESBL, tudi takrat, ko se beseda "ESBL" ne pojavlja v izvidih. Zato smo pred učenjem naših modelov in klasifikacijo iz vseh izvidov odstranili besedo "ESBL". Pravilnost smo preverjali s pomočjo podatkov, pridobljenih na Inštitutu za mikrobiologijo in imunologijo. Na inštitutu prisotnost bakterije *E. coli*, pozitivne na ESBL, preverijo z laboratorijsko preiskavo (opisano v poglavju 2.1.1).

5.1 Rezultati klasifikatorjev

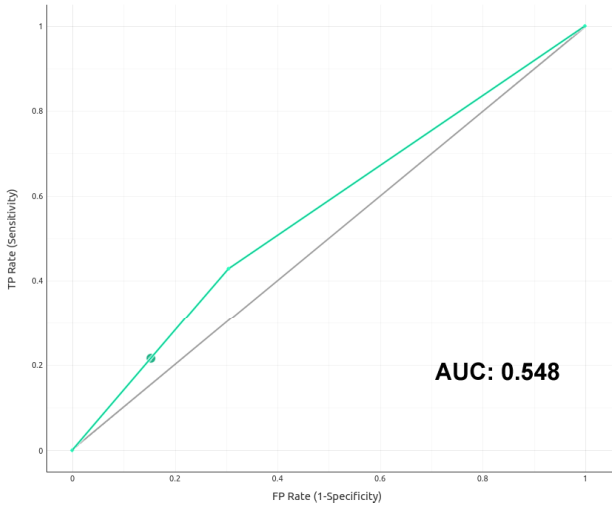
Predstavitev rezultatov posameznih klasifikatorjev s pomočjo krivulje ROC, metriko AUC in tabelo klasifikacij (angl. confusion matrix). Tabela klasifikacij se uporablja za predstavitev rezultatov dvo-ali večrazrednega klasifikatorja. Stolpci predstavljajo napovedi, vrstice pa pravilne vrednosti. Pravilne napovedi so tiste, kjer se pravilna in napovedana vrednost ujemata (primer tabele klasifikacij predstavlja slika 5.1).



Slika 5.1: Tabela klasifikacij, kjer so polja s pravilnimi napovedmi pobarvana zeleno ter polja z napačnimi napovedmi pobarvana rdeče (številke predstavljajo rezultat modela zdravnika) [39].

5.1.1 SGD

Predstavitev rezultatov klasifikatorja SGD s pomočjo krivulje ROC, oceno AUC (slika 5.2) in tabelo klasifikacij (tabela 5.1). Od 199 primero, je klasifikator pravilno napovedal 156 napovedi, od tega 151, kjer je ESBL enak 0, in 5, kjer je ESBL enak 1.



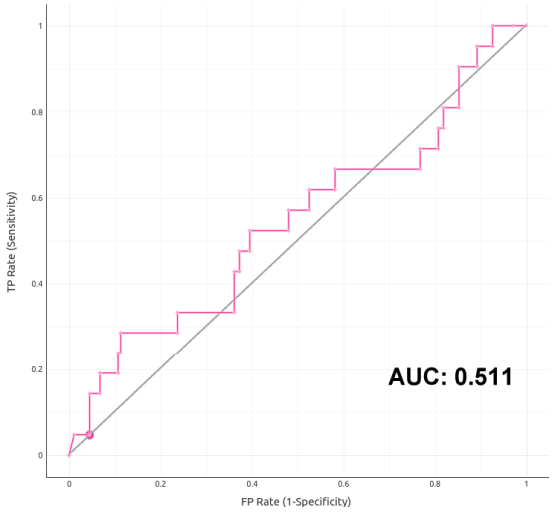
Slika 5.2: Krivulja ROC klasifikatorja SGD z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.1.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	151	27	178
	Je ESBL	16	5	21
	∑	167	32	199

Tabela 5.1: Tabela klasifikacij rezultatov testiranja klasifikatorja SGD.

5.1.2 Nevronska mreža

Predstavitev rezultatov nevronske mreže s pomočjo krivulje ROC, oceno AUC (slika 5.3) in tabelo klasifikacij (tabela 5.2). Od 199 primerov je klasifikator pravilno napovedal 171 napovedi, od tega 170, kjer je ESBL enak 0, in 1 napoved, kjer je ESBL enak 1.



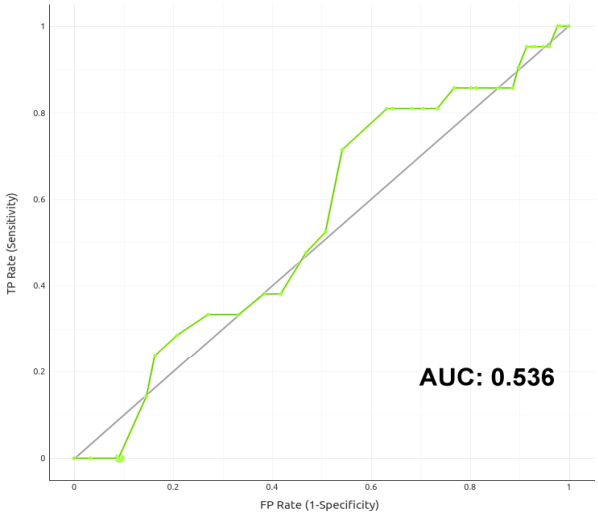
Slika 5.3: Krivulja ROC nevronske mreže z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.2.

		NAPOVED		
		Ni ESBL	Je ESBL	$\sum$
PRAVILNO	Ni ESBL	170	8	178
	Je ESBL	20	1	21
	$\sum$	190	9	199

Tabela 5.2: Tabela klasifikacij rezultatov testiranja nevronske mreže.

5.1.3 CN2 klasifikacijska pravila (entropy)

Predstavitev rezultatov klasifikacijski pravil CN2 (uporabljena ocenjevalna funkcija: entropy) s pomočjo krivulje ROC, oceno AUC (slika 5.4) in tabelo klasifikacij (tabela 5.3). Od 199 primerov je klasifikator pravilno napovedal 162 napovedi, od tega 162, kjer je ESBL enak 0, in 0, kjer je ESBL enak 1.



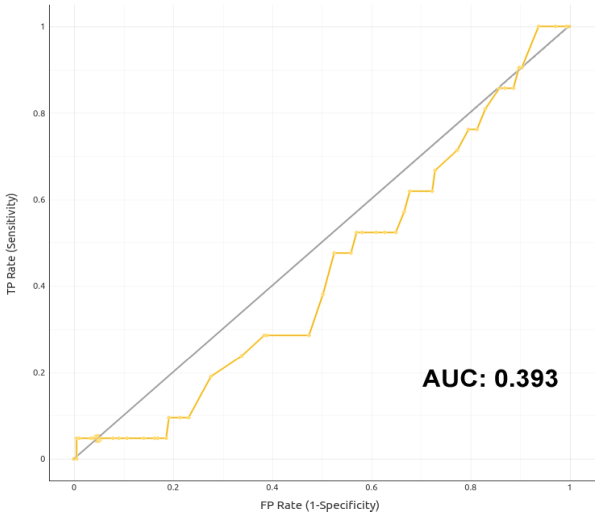
Slika 5.4: Krivulja ROC klasifikacijskih pravil CN2 z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.3.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	162	16	178
	Je ESBL	21	0	21
	∑	183	16	199

Tabela 5.3: Tabela klasifikacij rezultatov testiranja klasifikacijskih pravil CN2.

5.1.4 CN2 klasifikacijska pravila (Laplace)

Predstavitev rezultatov klasifikacijski pravil CN2 (uporabljena ocenjevalna funkcija: Laplace) s pomočjo krivulje ROC, oceno AUC (slika 5.5) in tabelo klasifikacij (tabela 5.4). Od 199 primerov je klasifikator pravilno napovedal 172 napovedi, od tega 171, kjer je ESBL enak 0, in 1 napoved, kjer je ESBL enak 1.



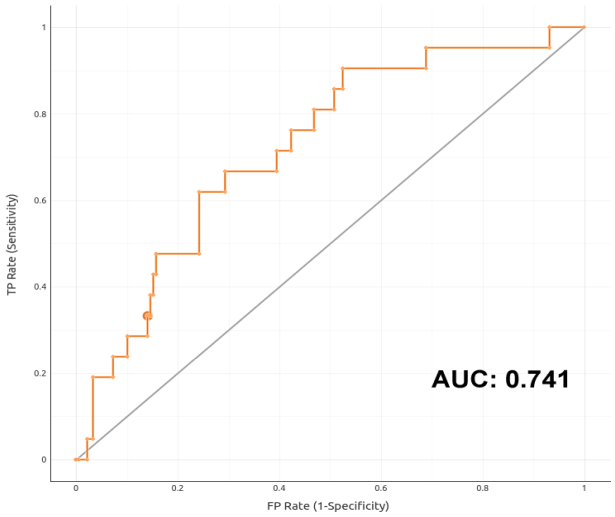
Slika 5.5: Krivulja ROC klasifikacijskih pravil CN2 z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.4.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	171	7	178
	Je ESBL	20	1	21
	∑	191	8	199

Tabela 5.4: Tabela klasifikacij rezultatov testiranja klasifikacijskih pravil CN2.

5.1.5 Naivni Bayesov klasifikator

Predstavitev rezultatov naivnega Bayesovega klasifikatorja s pomočjo krivulje ROC, oceno AUC (slika 5.6) in tabelo klasifikacij (tabela 5.5). Od 199 primerov je klasifikator pravilno napovedal 160 napovedi, od tega 153, kjer je ESBL enak 0, in 7, kjer je ESBL enak 1.



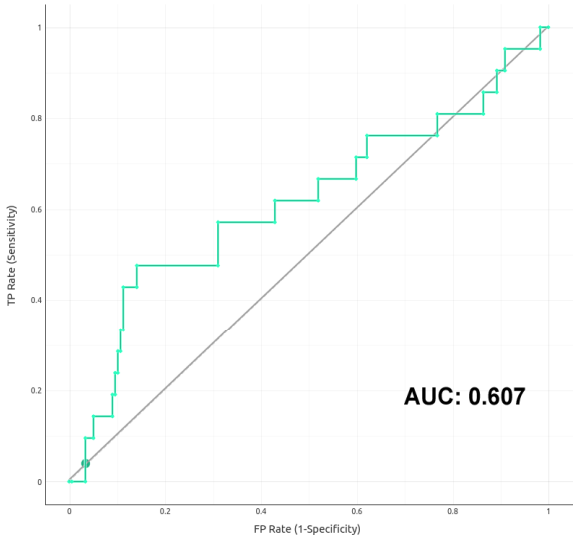
Slika 5.6: Krivulja ROC naivnega Bayesovega klasifikatorja z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.5.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	153	25	178
	Je ESBL	14	7	21
	∑	167	32	199

Tabela 5.5: Tabela klasifikacij rezultatov testiranja naivnega Bayesovega klasifikatorja.

5.1.6 Logistična regresija (L1)

Predstavitev rezultatov logistične regresije (z uporabo regularizacije L1) s pomočjo krivulje ROC, oceno AUC (slika 5.7) in tabelo klasifikacij (tabela 5.6). Od 199 primerov je klasifikator pravilno napovedal 173 napovedi, od tega 172, kjer je ESBL enak 0, in 1 napoved, kjer je ESBL enak 1.



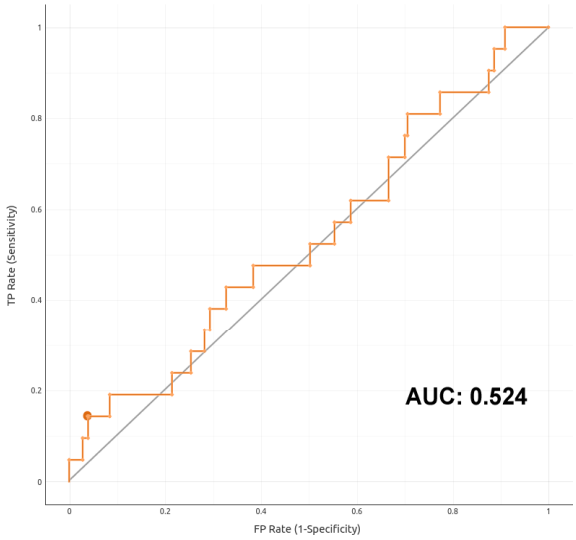
Slika 5.7: Krivulja ROC logistične regresije z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.6.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	172	6	178
	Je ESBL	20	1	21
	∑	192	7	199

Tabela 5.6: Tabela klasifikacij rezultatov testiranja logistične regresije.

5.1.7 Logistična regresija (L2)

Predstavitev rezultatov logistične regresije (z uporabo regularizacije L1) s pomočjo krivulje ROC, oceno AUC (slika 5.8) in tabelo klasifikacij (tabela 5.7). Od 199 primerov je klasifikator pravilno napovedal 172 napovedi, od tega 169, kjer je ESBL enak 0, in 3, kjer je ESBL enak 1.



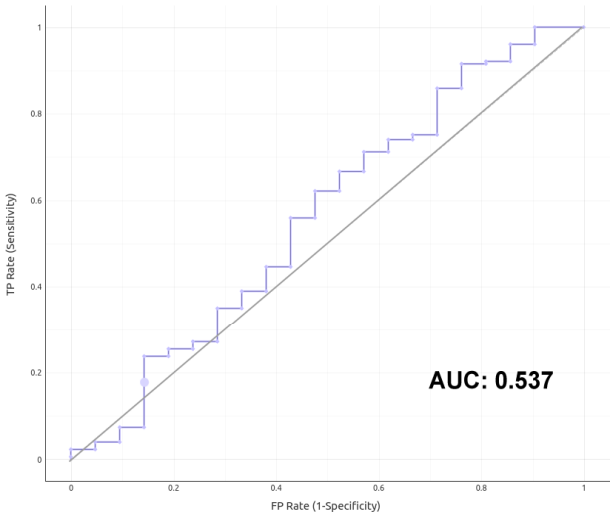
Slika 5.8: Krivulja ROC logistične regresije z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.7.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	169	9	178
	Je ESBL	18	3	21
	∑	187	12	199

Tabela 5.7: Tabela klasifikacij rezultatov testiranja logistične regresije.

5.1.8 SVM (linearno jedro)

Predstavitev rezultatov metode podpornih vektorjev s pomočjo krivulje ROC, oceno AUC (slika 5.9) in tabelo klasifikacij (tabela 5.8). Od 199 primerov je klasifikator pravilno napovedal 159 napovedi, od tega 155, kjer je ESBL enak 0, in 4, kjer je ESBL enak 1.



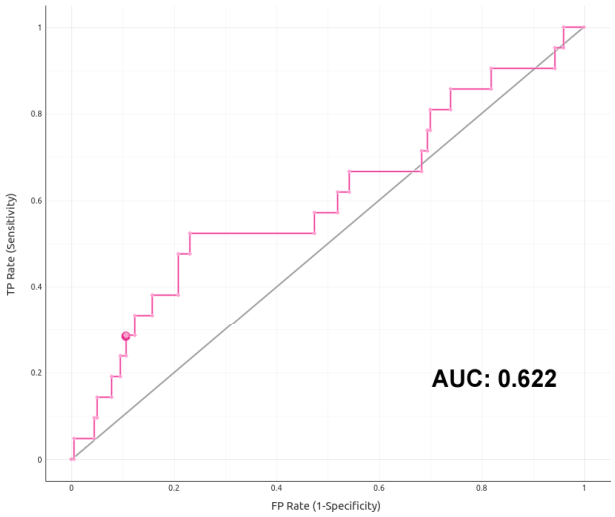
Slika 5.9: Krivulja ROC metode podpornih vektorjev z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.8.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	155	23	178
	Je ESBL	17	4	21
	∑	172	27	199

Tabela 5.8: Tabela klasifikacij rezultatov testiranja metode podpornih vektorjev.

5.1.9 SVM (sigmoidna jedrna funkcija)

Predstavitev rezultatov metode podpornih vektorjev s pomočjo krivulje ROC, oceno AUC (slika 5.10) in tabelo klasifikacij (tabela 5.9). Od 199 primerov je klasifikator pravilno napovedal 166 napovedi, od tega 160, kjer je ESBL enak 0, in 6, kjer je ESBL enak 1.



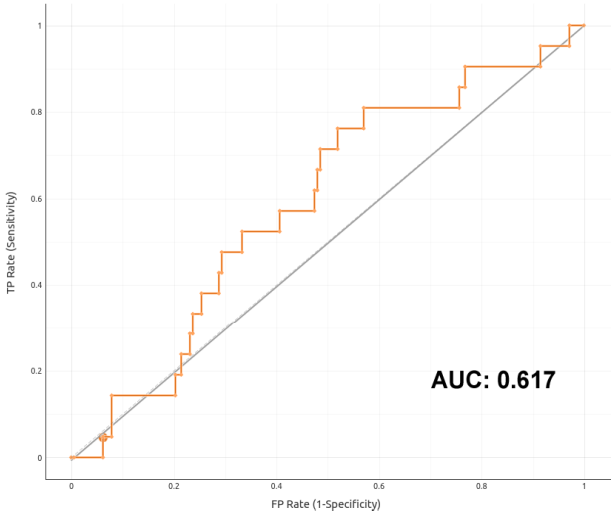
Slika 5.10: Krivulja ROC metode podpornih vektorjev z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.9.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	160	18	178
	Je ESBL	15	6	21
	∑	175	24	199

Tabela 5.9: Tabela klasifikacij rezultatov testiranja metode podpornih vektorjev.

5.1.10 SVM (jedrna funkcija RBF)

Predstavitev rezultatov metode podpornih vektorjev s pomočjo krivulje ROC, oceno AUC (slika 5.11) in tabelo klasifikacij (tabela 5.10). Od 199 primerov je klasifikator pravilno napovedal 170 napovedi, od tega 169, kjer je ESBL enak 0, in 1 napoved, kjer je ESBL enak 1.



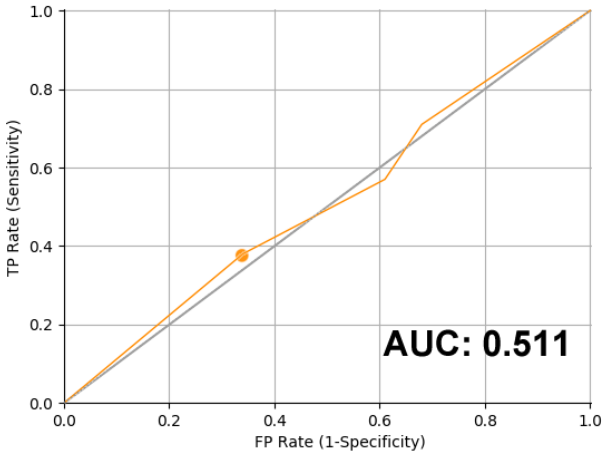
Slika 5.11: Krivulja ROC metode podpornih vektorjev z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.10.

		NAPOVED		
		Ni ESBL	Je ESBL	$\sum$
PRAVILNO	Ni ESBL	169	9	178
	Je ESBL	20	1	21
	$\sum$	189	10	199

Tabela 5.10: Tabela klasifikacij rezultatov testiranja metode podpornih vektorjev.

5.1.11 Uteženi multinomialni naivni Bayesov klasifikator

Predstavitev rezultatov metode podpornih vektorjev s pomočjo krivulje ROC, oceno AUC (slika 5.12) in tabelo klasifikacij (tabela 5.11). Od 199 primerov je klasifikator pravilno napovedal 124 napovedi, od tega 116, kjer je ESBL enak 0, in 8, kjer je ESBL enak 1.



Slika 5.12: Krivulja ROC uteženega multinomialnega naivnega Bayesovega klasifikatorja z oceno AUC. Točka na krivulji predstavlja rezultate iz tabele 5.11.

		NAPOVED		
		Ni ESBL	Je ESBL	∑
PRAVILNO	Ni ESBL	116	62	178
	Je ESBL	13	8	21
	∑	129	70	199

Tabela 5.11: Tabela klasifikacij rezultatov testiranja uteženega multinomialnega naivnega Bayesovega klasifikatorja.

5.2 Primerjava rezultatov

Rezultate smo zbrali in jih postavili v tabelo. Posamezne metode so postavljene v vrstice, metrike za ocenjevanje pa v stolpce. Krepki tisk predstavlja najboljši rezultat v stolpcu. Za boljšo primerjavo med samimi metodami smo ustvarili še grafični prikaz (slika 5.13).

Metoda	Ocena F1	Preciznost	Senzitivnost	Specifičnost	Klasifikacijska točnost
SGD	0.189	0.156	0.238	0.848	0.784
Nevronska Mreža	0.067	0.111	0.048	0.955	0.859
CN2 (entropy)	0.000	0.000	0.000	0.910	0.814
CN2 (Laplace)	0.069	0.125	0.048	0.910	0.864
Naive Bayes	0.264	0.219	**0.333**	0.860	0.804
Logistična regresija L1	0.071	0.143	0.048	**0.966**	**0.869**
Logistična regresija L2	0.182	**0.250**	0.143	0.949	0.864
SVM (linear)	0.167	0.148	0.190	0.871	0.799
SVM (sigmoid)	**0.267**	**0.250**	0.286	0.899	0.834
SVM (RBF)	0.065	0.100	0.048	0.949	0.854

Tabela 5.12: Rezultati metod uporabljenih v orodju Orange.

Metoda	Ocena F1	Preciznost	Senzitivnost	Specifičnost	Klasifikacijska točnost
Uteženi multinomialni naivni Bayesov klasifikator	0.176	0.114	0.380	0.652	0.623

Tabela 5.13: Rezultati uteženega multinomialnega naivnega Bayesa.

Metoda	Ocena F1	Preciznost	Senzitivnost	Specifičnost	Klasifikacijska točnost
Model zdravnika	0.440	0.450	0.430	0.940	0.880

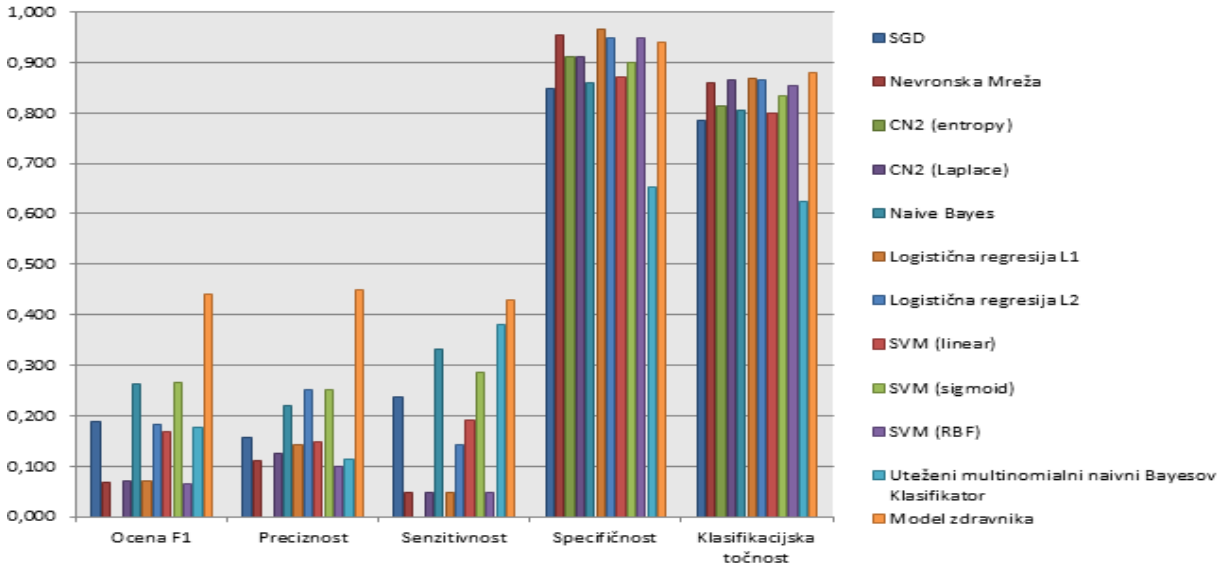
Tabela 5.14: Rezultati modela zdravnika.

	Model zdravnika		SVM (sigmoid)
Lažno pozitiven	**11**	<	18
Lažno negativen	**12**	<	15

Tabela 5.15: Primerjava napačnih napovedi modela zdravnika in metode podpornih vektorjev (z uporabo sigmoidne jedrne funkcije). Primerjava napovedi, ko je bil napovedan ESBL, in ne bi smel biti (lažno pozitiven) in ko ni bil napovedan, pa bi moral biti (lažno negativen).

	Model zdravnika		Naivni Bayesov klasifikator
Lažno pozitiven	**11**	<	25
Lažno negativen	**12**	<	14

Tabela 5.16: Primerjava napačnih napovedi modela zdravnika in naivnega Bayesovega klasifikatorja. Primerjava napovedi, ko je bil napovedan ESBL, in ne bi smel biti (lažno pozitiven) in ko ni bil napovedan, pa bi moral biti (lažno negativen).



Slika 5.13: Primerjava vseh metod po metrikah, ki smo jih uporabili za merjenje uspešnosti klasifikatorjev.

5.3 Kvantitativna analiza

Za primerjavo rezultatov smo izbrali najboljši metodi (ki smo ju uporabili) in ju primerjali z modelom zdravnika. V tabeli 5.17 so prikazane primerjave med modelov zdravnika in metodo podpornih vektorjev (z uporabo sigmoidne jedrne funkcije), v tabeli 5.18 pa so prikazane primerjave med modelom zdravnika in naivnim Bayesovim klasifikatorjem. Obe metodi, metoda podpornih vektorjev in naivni Bayesov klasifikator, sta metodi iz orodja Orange.

	Model zdravnika	SVM (sigmoid)	[1] Razlika
Ocena F1	0.450	0.267	-0.183 (-40.6%)
Preciznost	0.400	0.250	-0.150 (-37.5%)
Senzitivnost	0.450	0.286	-0.164 (-36.4%)
Specifičnost	0.940	0.899	-0.041 (-4.4%)
Klasifikacijska točnost	0.890	0.834	-0.056 (-6.3%)

Tabela 5.17: Primerjava rezultatov modela zdravnika z metodo podpornih vektorjev (s sigmoidno jedrno funkcijo).

	Model zdravnika	Naivni Bayesov klasifikator	[1] Razlika
Ocena F1	0.450	0.264	-0.186 (-41.3%)
Preciznost	0.400	0.219	-0.181 (-45.3%)
Senzitivnost	0.450	0.333	-0.117 (-25.6%)
Specifičnost	0.940	0.860	-0.080 (-8.5%)
Klasifikacijska točnost	0.890	0.804	-0.086 (-9.7%)

Tabela 5.18: Primerjava rezultatov modela zdravnika z naivnim Bayesovim klasifikatorjem (iz orodja Orange).

[1]Razlika modela v primerjavi z modelom zdravnika.

5.4 Kvalitativna analiza

Problem, ki ga poskušamo rešiti, je specifičen. Senzitivnost, ki smo jo dosegli, je nižja, kot če bi se odločali o tem zgolj na podlagi podatkov o koloniziranosti bolnika z bakterijami, pozitivnimi na encime ESBL. Za potencialno uporabo bi morala biti senzitivnost vsaj 65% in še v tem primeru bi bila uporabna zgolj za ogrožene bolnike. Specifičnost (90%) je zadovoljiva. Klasifikator zaradi prenizke senzitivnosti (28% za SVM oz. 33% za naivni Bayesov klasifikator) ne more biti uporabljen kljub visoki specifičnosti (90% za SVM oz. 86% za naivni Bayesov klasifikator). Zaradi občutljivosti problema (vse večja odpornost bakterij proti protimikrobnim sredstvom) moramo imeti visoko specifičnost kot tudi senzitivnost.

Poglavje 6

Zaključek

V delu smo spoznali problem odpornosti mikroorganizmov proti protimokrobnim sredstvom. Osredotočili smo se zgolj na odpornost mikroorganizmov proti protimikrobnim sredstvom iz razreda betalaktamov. V našem primeru je bila to bakterija *Escherichia coli*, ki proizvaja encim - beta laktamazo. Število mikroorganizmov, odpornih proti protimikrobnim sredstvom narašča. Zadnji razred protimikrobnih sredstev pa je bil odkrit leta 1987.

Predpisovanje napačnih protimikrobnih sredstev (npr. predpisovanje protimikrobnih sredstev, proti katerim je bakterija, s katero je bolnik okužen, odporna) zgolj še povečuje odpornost mikroorganizmov proti protimikrobnim sredstvom. Problem smo poskušali rešiti s pomočjo strojnega učenja in tekstovnega rudarjenja. Iz zdravniškega izvida smo poskušali napovedati, ali ima bolnik v urinu prisotno bakterijo *Escherichio coli*, ki proizvaja encime ESBL.

Med zbiranjem izvidov smo naleteli na težavo, in sicer na zelo zamudno zbiranje izvidov. Najprej je bilo treba pridobiti seznam bolnikov, ki nam ga je predal Inštitut za mikrobiologijo in imunologijo. Ključnega pomena je bilo, da so na teh bolnikih izvedli laboratorijsko preiskavo, s katero so preverili, ali je bolnik okužen z bakterijo, pozitivno na encime ESBL, ali ne. Ti podatki so nam namreč služili za preverjanje naših rezultatov. S pomočjo tega seznama smo nato na Kliniki za infekcijske bolezni in vročinska stanja zbirali nadaljnje podatke. Seznam bolnikov je v papirnati obliki, tako da je bilo treba imena in priimke bolnikov ročno prepisovati v program na računalniku, od koder smo nato izvid za izvidom prekopirali in anonimizirali. Vsaj za zdaj je informacijska organiziranost dokaj slaba in je raziskovalno delo zelo oteženo. Kljub informacijskim sistemom je za raziskovalno delo, ki zahteva več podatkov o večih bolnikih, treba čakati na drugo ustanovo, da pošlje podatke, ti pa so v papirnati obliki.

Besedila (izvide, napisane v naravnem jeziku) smo predhodno obdelali (korenjenje

besed in filtriranje besed z določenimi značkami). Iz besedil smo odstranili tudi besedo "ESBL", saj se zdravniki iz izvidov, kjer koloniziranost z ESBL ni omenjena, težje odločijo za pravilno izbiro protimikrobnega sredstva. Izvide smo tudi analizirali. S pomočjo metode ANOVA smo poiskali besede z največjo težo. Zaradi potreb orodja Orange, smo podatke preoblikovali v matriko, kjer so stolpci predstavljali frekvence posameznih preoblikovanih (korenjenje,...) besed, vrstice pa besedila (izvide). Preizkusili smo več napovednih modelov. Za najboljšega se je izkazala metoda podpornih vektorjev iz orodja Orange. Čeprav smo dosegli visoko specifičnost (90% za SVM oz. 86% za naivni Bayesov klasifikator), klasifikator zaradi prenizke senzitivnosti (28% za SVM oz. 33% za naivni Bayesov klasifikator) in občutljivosti problema ne more biti uporabljen. Problem vse večje odpornosti bakterij proti protimikrobnim sredstvom zahteva, da imamo visoko specifičnost, kot tudi senzitivnost.

Če izboljšamo model, bi naš model lahko bil uporaben že pri samem sprejemu bolnikov. Po vnesenih podatkih (anamneza in status) o bolniku bi sistem avtomatsko analiziral izvid in zdravniku namignil, da ima bolnik lahko okužbo z bakterijo, ki proizvaja encime ESBL.

6.1 Kritična analiza

Iz besedila (izvidov) smo odstranili besede "ESBL", da bi ugotovili, kako dobro deluje klasifikator na primerih, ki so za zdravnike težji. Rezultati bi lahko bili boljši tudi, če bi imeli na voljo več izvidov, vendar smo zaradi omejitve s časom in zamudnega zbiranja izvidov za magistrsko nalogo zbrali zgolj 210 izvidov.

6.2 Ideje za izboljšave in nadaljnje delo

Rezultate bi lahko izboljšali z večjo množico izvidov. Zaradi omejitve s časom smo preizkusili zgolj peščico modelov. V nadaljevanju nameravamo preizkusiti še druge metode (npr. Word2Vec [40]), kombiniranje modelov, ansambelske metode, poskus z nadvzorčenjem manjšinskega razreda (v našem primeru, ko je ESBL enak 1) in predvsem povečati število izvidov. Vredno bi bilo poskusiti tudi razbrati vrednosti laboratorijskih preiskav (če so prisotne v izvidu) in jih dodati med atribute poleg frekvenc besed.

Literatura

[1] Rutger Hermsen, J. Barret Deris, Terence Hwa *On the rapidity of antibiotic resistance evolution facilitated by a concentration gradient.*, Proceedings of the National Academy of Sciences of the United States of America 109:10775-10780, maj 2012.

[2] Laura A. Dever, Terence S. Dermody *Mechanisms of Bacterial Resistance to Antibiotics*, JAMA Internal Medicine, pp. 886-895, American Medical Association, maj 1991.

[3] Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain *Natural Language Processing*, International Journal of Technology Enhancements and Emerging Engineering Research, pp. 131-134, IJTEEE 2016.

[4] Public Health England *Extended-spectrum beta-lactamases (ESBLs*: treatment, prevention, surveillance), Extended-spectrum beta-lactamases (ESBLs): guidance, data, analysis, december 2013.

[5] David L. Paterson, Robert A. Bonomo *Extended-spectrum beta-lactamases: a clinical update.*, Clinical Microbiology Reviews, American Society for Microbiology (ASM), 2005

[6] Shakti Rath, Debasmita Dubey, Mahesh C. Sahu, Rabindra N Padhy *Surveillance of ESBL producing multidrug resistant Escherichia coli in a teaching hospital in India*, Asian Pacific Journal of Tropical Disease, pp. 140-149, Hainan Medical College 2014.

[7] Harris AD, McGregor JC, Johnson JA, Strauss SM, Moore AC, et al. *Risk factors for colonization with extended-spectrum beta-lactamase-producing bacteria and intensive care unit admission.*, Emerging Infectious Diseases, pp. 1144-1149, Centers for Disease Control and Prevention (CDC), avgust 2007.

[8] Lynn L. Silver, *Challenges of Antibacterial Discovery*, Clinical Microbiology Reviews, pp. 71-109, American Society for Microbiology (ASM), januar 2011.

[9] Stanford T. Shulman, Herbert C. Friedmann, Ronald H. Sims *Theodor Escherich: The First Pediatric Infectious Diseases Physician?*, Clinical Infectious Diseases, pp. 1025–1029, Oxford academic, oktober 2007.

[10] Ji Youn Lim, Jang W. Yoon, Carolyn J. Hovde *A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O157*, Journal of Microbiology and Biotechnology, pp. 5-14, Springer, januar 2010.

[11] Aravind J. Joshi *Natural Language Processing*, American Association for the Advancement of Science, september 1991.

[12] Igor Kononenko, Matjaž Kukar. Support vector machines. *Machine Learning and Data Mining*, pp 267-273, Woodhead Publishing, 2007.

[13] Igor Kononenko, Matjaž Kukar. Perceptron. *Machine Learning and Data Mining*, pp 298-306, Woodhead Publishing, 2007.

[14] Igor Kononenko, Matjaž Kukar. Logistic regression. *Machine Learning and Data Mining*, pp 267-267, Woodhead Publishing, 2007.

[15] Igor Kononenko, Matjaž Kukar. Naive and semi-naive Bayesian classifier. *Machine Learning and Data Mining*, pp 242-249, Woodhead Publishing, 2007.

[16] Léon Buttou. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010, 19th International Conference on Computational StatisticsParis France*, pp 177-186, Springer, 2010.

[17] Peter Clark, Tim Niblett. The CN2 induction algorithm. In *Machine Learning*, pp 261–283, Springer, 1988.

[18] Bernhard Scholkopf, Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press Cambridge, MA, USA, 2001

[19] Jorge Nocedal, Stephen J. Wright *Numerical Optimization, Second edition*, Springer, 2006.

[20] Pete Chapman, Julian Clinton, Randy Kerber, et al. *CRISP-DM 1.0, Step-by-step data mining guide.* The CRISP-DM consortium, August 2000.

[21] Igor Kononenko, Marko Robnik Šikonja. Krivulja ROC. *Inteligentni sistemi*, pp 63-64, Založba FE in FRI, 2010.

[22] ANOVA Analysis of Variance. Introduction to Statistics for Biology and Biostatistics by Susan Holmes, Stanford University, 2004.
http://statweb.stanford.edu/~susan/courses/s141/hoanova.pdf. Zadnji dostop 22.2.2018.

[23] One-Way ANOVA. Introduction to Statistics by James Jones, Richland Community College.
https://people.richland.edu/james/lecture/m170/ch13-1wy.html. Zadnji dostop 22.2.2018.

[24] SVM - jedrne metode (Support Vector Machines & Kernels). Introduction to Machine Learning by Eric Eaton, Ph. D., University of Pennsylvania.
http://www.seas.upenn.edu/~cis519/fall2017/. Zadnji dostop 18.2.2018.

[25] Dokumentacija za orodje ReLDI.
http://nl.ijs.si/ME/V5/msd/html/msd-sl.html. Zadnji dostop 15.2.2018.

[26] Classification and logistic regression. *CS229 Lecture notes, by Andrew Ng.* http://cs229.stanford.edu/notes/cs229-notes1.pdf. Zadnji dostop 25. januar 2018.

[27] Andrew Y. Ng *Feature selection, L1 vs. L2 regularization, and rotational invariance.* pp 78-85. Proceeding ICML '04 Proceedings of the twenty-first international conference on Machine learning, 2004.

[28] Multinomial naive Bayesian classifier https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html. Zadnji dostop 23. januar 2018.

[29] National Institute of Allergy and Infectious Diseases. https://www.niaid.nih.gov/sites/default/files/E.coli_.jpg. Zadnji dostop 12. december 2017.

[30] Healthline. https://www.healthline.com/health/esbl#esbl-transmission. Zadnji dostop 12. december 2017.

[31] World Health Organization (WHO). http://www.who.int/mediacentre/factsheets/fs194/en. Zadnji dostop 4. december 2016.

[32] Orange, Data Mining tool (University of Ljubljana, Faculty of Computer and Information Science, Biolab) https://orange.biolab.si. Zadnji dostop 15. november 2017.

[33] Sigmoidna funkcija. http://mathworld.wolfram.com/SigmoidFunction.html. Zadnji dostop 13. december 2017.

[34] NLTK, Natural Language Toolkit (NLTK Project) http://www.nltk.org. Zadnji dostop 15. november 2017.

[35] Snowball, Snowball stem https://snowballstem.org. Zadnji dostop 15. november 2017.

[36] ReLDI, ReLDI tagger
https://reldi.spur.uzh.ch/blog/tagger/. Zadnji dostop 15. november 2017.

[37] ReLDI-tagger & lemmatiser
https://github.com/clarinsi/reldi-tagger. Zadnji dostop 15. november 2017.

[38] NLTK - Plaintext Corpus Reader
http://www.nltk.org/_modules/nltk/corpus/reader/plaintext.html. Zadnji dostop 15. november 2017.

[39] Simple guide to confusion matrix terminology
http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. Zadnji dostop 19. januar 2018.

[40] Word2Vec
https://deeplearning4j.org/word2vec.html. Zadnji dostop 9. januar 2018.

[41] TextToMatrix - Orodje za pretvorbo besedila v matriko (Sandi Mikuš).
https://github.com/smiks/tools/blob/master/TextToMatrix.py. Zadnji dostop 15. november 2017.