Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Jernej Henigman

Strojno učenje kemijskih reakcij proteinov v interakciji z RNA

diplomsko delo

univerzitetni študijski program prve stopnje
računalništvo in informatika

Mentor: doc. dr. Tomaž Curk

Ljubljana 2015

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

V diplomski nalogi preučite kemoinformacijske pristope za opisovanje strukturnih sprememb molekul v kemijskih reakcijah v celici. Za izbrane skupine kemijskih reakcij določite najbolj ustrezen pristop modeliranja s strojnim učenjem. Tako določen pristop uporabite za izgradnjo napovednega modela kemijskih reakcij, ki nastajajo v kontekstu encimov, ki imajo sposobnost vezave na RNA. Poročajte o uspešnosti napovednega modela ter o najpomembnejših strukturnih spremembah v kemijskih reakcijah prej omenjenih encimov.

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Jernej Henigman, z vpisno številko **63100259**, sem avtor diplomskega dela z naslovom:

Strojno učenje kemijskih reakcij proteinov v interakciji z RNA

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom doc. dr. Tomaža Curka,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 17.8.2015 Podpis avtorja:

Zahvaljujem se očetu Benjaminu ter mami Nevenki za moralno podporo med izdelavo diplomske naloge. Posebej pa bi se rad zahvalil mentorju dr. Tomažu Curku za vso ponujeno pomoč, za vse nasvete, vse tedenske sestanke, za vsak odgovorjen e-mail, vsak komentar, za potrpežljivost, kolegialnost ter prijaznost.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
RNA	ribonucleic acid	ribonukleinska kislina
EC	enzyme commission	skupina encimov
SVM	support vector machine	metoda podpornih vektorjev
SGD	stohastic gradient descent	stohastični gradientni spust
DT	decision tree	odločitveno drevo
NB	naive Bayes	naivni Bayes
RF	random forest	naključni gozd
NN	neural networks	nevronske mreže
AUC	area under the ROC curve	površina pod krivuljo ROC

Povzetek

V diplomski nalogi smo uporabili metode strojnega učenja za klasifikacijo kemijskih reakcij. Hkrati smo identificirali najpomembnejše strukturne spremembe molekul, ki nastajajo v kemijskih reakcijah proteinov, ki sicer vstopajo tudi v interakcijo z RNA. V prvem delu smo na podlagi šestih osnovnih encimatskih skupin kemijskih reakcij, določili optimalen nabor parametrov modeliranja. Testirali smo tri skupine parametrov: metode za uravnoteževanje učne množice (sedem metod), metode za opisovanje kemijskih profilov (sedem vrst opisov) in metode za izgradnjo napovednih modelov (pet metod). Za najboljši nabor smo izbrali kombinacijo tistih parametrov, pri kateri je mera AUC najvišja. Empirično smo pokazali, da najboljši nabor sestavlja kombinacija parametrov: uravnoteževalna metoda naključnega podvzorčenja, združena kemijska profila Morgan in MorganBitVector, napovedni model naključnega gozda. Dosežen povprečen AUC na šestih osnovnih skupinah kemijskih reakcij je znašal 0,97. V drugem delu smo uporabili predhodno določen nabor parametrov za modeliranje skupine kemijskih reakcij proteinov v interakciji z RNA, pri čemer smo dosegli AUC 0,77.

Ključne besede: strojno učenje, kemijske reakcije, uravnoteževalne metode, kemijski opisi molekul, napovedni modeli, AUC, RNA, struktura molekule.

Abstract

In this thesis machine learning methods are used to classify chemical reactions. At the same time the most important changes in molecular structure are identified that are typical for chemical reactions of RNA-binding proteins. In the first part, six basic groups of chemical reactions were used to determine the optimal set of parameters for modeling and prediction. Three groups of parameter sets were tested: methods for balancing the learning set (seven methods), methods for molecular fingerprinting (seven methods) and predictive models (five methods). Empirically is shown that the best combination consists of the following parameters: random undersampling as balancing method, Morgan+MorganBitVector for molecular fingerprinting and random forest as predictive model, with which average AUC 0.97 was achieved. For the second part, the optimal set of parameters is used to discriminate between chemical reactions associated with RNA-binding proteins and those chemical reactions associated with non RNA-binding proteins. AUC score 0.77 was achieved.

Keywords: machine learning, chemical reactions, balancing methods, molecular fingerprints, predictive models, AUC, RNA, molecular structure.

Poglavje 1

Uvod

Motivacija za izdelavo diplomske naloge izhaja iz vprašanja, kako metabolizem vpliva na reguliranje genov (angl. gene regulation) in obratno. Prof. Matthias Hentze je podal naslednji primer [18]: "Predstavljajte si, da se vzpenjate na osemtisočaka v Himalaji, kjer ste podvrženi hudemu fizičnemu naporu in pomanjkanju kisika. Čez 14 dni pa že brezskrbno poležavate na sončni plaži Maldivov. Če bi primerjali izražanje genov (angl. gene expression) in metabolizem pod tema dvema različnima pogojema, bi med njima opazili izrazito razliko. Vprašanje, ki se poraja na tem mestu, je, kako so med seboj povezani encimi, metabolizem in mehanizmi uravnavanja izražanja genov". Trenutne znanstvene raziskave v splošnem odgovor na to vprašanje iščejo v povezavi s proteini, ki opravljajo dvojno funkcijo (angl. moonlighting proteins). Ti proteini imajo zanimivo lastnost. Skozi evolucijo so se razvili tako, da so začeli opravljati dodatno nalogo poleg že obstoječe. Zanimajo nas proteini, ki kot encimi nastopajo v metaboličnih kemijskih reakcijah in hkrati vstopajo v interakcijo z RNA. Širši cilj raziskav tega še precej neraziskanega področja je dokazati, da se evkariotske celice prilagajajo na spremembe metaboličnih pogojev, kot so razpoložljivost hranil, vsebnost kisika in vpliv stresa, post-translacijsko in ne preko mehanizma za uravnavanje izražanja genov, ki trenutno velja za splošno sprejet vidik [19].

1.1 Modeliranje kemijskih reakcij

Do zdaj je bilo izvedenih veliko projektov v povezavi z genomskimi zaporedji različnih živih organizmov, kjer so generirali zelo veliko podatkov o genih in proteinih. Ker so eksperimentalni pristopi karakterizacije funkcijskih razredov proteinov dragi in časovno zahtevni, so se računalniške napovedne metode izkazale za dobro alternativo [5]. Mnoge metode strojnega učenja so bile uporabljene za napovedovanje ustrezne številke EC (angl. Enzyme Commission number) posameznih kemijskih reakcij. Številka EC je numerična klasifikacijska shema encimov. Temelji na vrsti kemijske reakcije, v kateri posamezen encim nastopa. Veliko reakcij, prisotnih v različnih bioloških poteh, nima ustrezne oznake EC zaradi pomanjkanja objavljenih člankov o encimatskih testih [5]. Številka EC predstavlja povezavo med informacijami genov in kemijskimi reakcijami, zato je ugotavljanje točnosti številk EC pomembna naloga. V grobem lahko metode za napovedovanje številk EC razdelimo na dva pristopa, na podlagi tipa uporabljene informacije za generiranje deskriptorjev: bioinformacijski in kemoinformacijski pristop.

1.1.1 Bioinformacijski pristop

Temelji na informacijah o proteinskih sekvencah aminokislin in strukturni zgradbi posameznega proteina [5]. Iz zaporedja proteina pridobimo potrebne informacije (atribute), na podlagi katerih klasificiramo protein v ustrezno skupino. Dozdajšnje delo je bilo usmerjeno v primerjavo sekvenčnih zapisov proteinov s pomočjo algoritmov za poravnavo sekvenc (angl. sequence alignement algorithms), kot so BLAST, PSI-BLAST, HMMER. Dobson [6], je razvil napovedni model za določanje številke EC na podlagi strukturne zgradbe proteina, kjer informacija o poravnavi sekvenc aminokislin ni bila več uporabljena. Dosegli so 60-odstotno točnost pri določevanju številk EC posameznih encimov. Bray [7] je združil obe metodi (poravnava sekvenc, strukturna zgradba proteina) in dosegel 16,4-odstotno izboljšavo točnosti glede na zgornjo raziskavo. Združena metoda je bila od naključnega algo-

ritma za določevanje encimatskega razreda boljša za 26,1%, kar je bil dovolj jasen kazalnik, da strukturna zgradba in zaporedja aminokislin proteina nosijo veliko informacij, s pomočjo katerih klasificiramo proteine v ustrezen razred EC. Kljub navidezni uspešnosti zgornjih raziskav se izkaže, da so globalni napovedni modeli za napovedovaje encimatskega razreda na podlagi proteinskih zgradb in sekvenc precej zapleteni, njihova točnost napovedovanja pa ne zadovolji pričakovanih rezultatov [5].

Naslednja od razvitih metod je metoda EzyPred. Razvita je bila kot skupek pristopov FunD (angl. Functional Domain) in Pse-PSSM (angl. Pseudo Position-Specific Scoring Matrices). EzyPred je triplastni prediktor. V prvi plasti algoritem poskuša identificirati, ali je opazovani protein encim ali ne. Druga in tretja plast pa poskušata napovedati pravilno številko EC encima. Metoda je bila testirana na podatkih iz baze ExplorerEnz. V podatkovni množici ni nobeden od proteinov vseboval več kot 40 % sekvenčne identitete kakšnega drugega proteina iz množice. Skupen dosežek treh plasti napovedovanja encimatskih številk je bil večji od 90 % [5]. EzyPred je prosto dostopna metoda na naslovu http://chou.med.harvard.edu/bioinf/EzyPred/. Implementacija ansambelskih tehnik, ki so se začele pojavljati kot posledica napredka na področju strojnega učenja, je predstavljala nov predor pri napovedni uspešnosti številk EC. Z uporabo metode podpornih vektorjev, pri čemer se upošteva polno hierarhična struktura številke EC, je možno napovedati pripadnost oksidoreduktaznim podskupinam s 93-odstotno točnostjo [5].

Pred kratkim je bil predstavljen nov napovedni model. Uporablja N-to-1 nevronsko mrežo. Metodo so testirali s pomočjo 10-kratnega prečnega preverjanja na veliki neredundantni označeni množici encimov, pridobljeni iz podatkovne baze UniProtKB. Izmerjena je bila 96-odstotna točnost klasifikacije. Razlog za visoko točnost rezultatov je v tem, da je N-to-1 nevronska mreža sposobna uporabe velikega števila prostih atributov, s pomočjo katerih zazna kompleksnejše in daljše vzorce ostankov, ki nosijo dodatno informacijo [5].

1.1.2 Kemoinformacijski pristop

Temelji na informacijah o kemijskih spremembah reakcij, v katerih nastopajo encimi. Od bioinformacijskega pristopa se razlikuje v tem, da informacije (atribute) za strojno učenje ne pridobi več neposredno iz proteinov (proteinske sekvence amino kislin, strukturna zgradba proteina), temveč posredno iz sprememb kemijskih reakcij, ki jih pospešujejo encimi [5]. Prvi trije nivoji številke EC opisujejo tvorjenje in razpadanje kemijskih vezi v reakciji, zato pri napovedovanju številke EC s kemoinformacijskim pristopom ne gre za napovedovanje, ampak določevanje številke EC posamezni kemijski reakciji, v kateri nastopa posamezen encim. Veliko raziskovalnih skupin se je ukvarjalo z računanjem in odkrivanjem novih kemijskih in topoloških deskriptorjev kemijskh reakcij. Poleg informaciji o spremembah kemijskih vezi obstajajo še informacije o reakcijskem vzorcu, naravi substrata, tipih prenosnih in sprejemnih skupin (angl. donor group, angl. acceptor group) [5]. Iz vseh teh lastnosti lahko pridobimo zelo veliko informacij, ki predstavljajo atribute za klasificiranje kemijskih reakcij v ustrezne razrede.

V raziskavi Latino and Aires-de-Sousa [10] uporabijo MOLMAP (angl. MOLecular Mapping of Atom-level Properties). To je reakcijski deskriptor, ki zakodira spremembe kemijskih vezi, ko poteče encimatska kemijska reakcija. MOLMAP deluje na principu Kohonenove nevronske mreže SOM (angl. Self-Organizing Map), ki definira tipe kovalentnih vezi na podlagi njihovih topoloških značilnosti. Z uporabo MOLMAP deskriptorejev (atributov) in uporabo strojne učne metode naključnega gozda (angl. random forest) so testirali uspešnost določevanja številk EC posameznim kemijskim reakcijam. 95, 90 in 85 % so bile točnosti za prvi, drugi in tretji nivo številke EC. Metoda za delovanje zahteva celotno kemijsko formulo.

Naslednja metoda za določevanje številke EC temelji na vzorcih RDM (angl. reaction center(R), the difference region(D), matched region(M)). RDM poskuša opisati vzorce strukturnih sprememb encimatskih kemijskih reakcij [11]. Za vsako kemijsko spremembo se tvorijo pari RDM, ki nosijo informacije o spremembah v reakciji. Metoda, ki smo jo povzeli po članku Ya-

manashi [12], deluje v treh korakih. V prvem koraku se izračuna vzorec RDM reakcije. V drugem koraku se primerja vzorec RDM iz prvega koraka z vzorci RDM že znanih in označenih številk EC. V tretjem koraku se izvede uteženo glasovanje za izbiro prave številke EC glede na podobnosti, izračunane v drugem koraku med vzorci RDM. S prečnim preverjanjem, je bilo potrjeno, da metoda dosega visoko točnost pri napovedovanju številk EC. Posledično je bil postavljen spletni strežnik, imenovan E-zyme, ki je prosto dostopen na naslovu http://www.genome.jp/tools/e-zyme/. Prednost sistema E-zyme je v tem, da za določevanje prave številke EC ne potrebuje celotne kemijske formule, temveč samo reaktantne pare, kar se izkaže za uporabno pri še nedoločenih reakcijah, saj te tipično ne vsebujejo celotne kemijske formule [5].

Do zdaj najboljša razvita metoda za določevanje številk EC je metoda ECOH (angl. Enzyme COmmission numbers Handler). Dostopna je na naslovu http://www.bioinfo.sk.ritsumei.ac.jp/apps/ecoh/. Uporablja algoritem MCS (angl. Maximum Common Substructure), ki temelji na strukturni podobnosti substratov in produktov posamezne encimatske reakcije. Algoritem je sestavljen iz treh korakov. V prvem koraku se izlušči značilne podstrukture iz substratov in produktov z uporabo algoritma MCS. V drugem se z uporabo metode medsebojne informacije izračuna podobnost med strukturami. V tretjem pa se uporabi metoda podpornih vektorjev za napoved številk EC. Izračunane točnosti prvih treh nivojev številk EC so 99,8, 87,4 in 83,7 %. Nekoliko slabše točnosti so se pokazale samo pri klasifikaciji izomernih reakcij [5].

1.2 Cilji

V splošnem želimo razviti metodo strojnega učenja za klasifikacijo kemijskih reakcij v izbrano skupino reakcij (binarni razred). Dodaten cilj diplomske naloge je poiskati pomembne stukturne spremembe, ki nastajajo kot produkt kemijskih reakcij, v katerih so udeleženi proteini, ki poleg encimatske funkcije vstopajo tudi v interakcije z RNA.

Poglavje 2

Podatki

Osnovni učni primer pri našem delu je kemijska reakcija. Kemijska reakcija opisuje molekule pred in po reakciji. Kemijske reakcije smo razdelili v ustrezne tarčne skupine. Posebej nas je zanimala skupina kemijskih reakcij, ki se dogajajo v kontekstu RNA-vezavnih proteinov. Podatke (kemijske reakcije, molekule, proteine) smo pridobili iz podatkovne baze KEGG.

2.1 Podatkovna baza KEGG

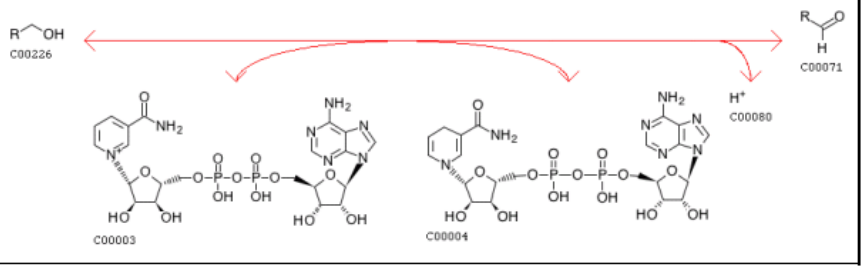
KEGG (angl. Kyoto Encyclopedia of Genes and Genomes) je zbirka podatkovnih baz v povezavi z genomi, biološkimi potmi, kemijskimi substancami in boleznimi. Gre za eno najbolj priljubljenih baz na področju izobraževanja, bioinformatike, kemije in drugih sorodnih ved. Ustvarjena je bila leta 1995 na univerzi v Kyotu pod vodstvom profesorja Minoruja Kanehisa. Od tedaj pa je bila posodobljena na vsaka dva meseca. Mnogi bazo KEGG poimenujejo računalniška reprezentacija biološkega sistema, saj povezuje osnovne biološke gradnike, kot so geni, proteini, molekule, reakcije, v smiselno organizirano celoto [34]. Slika 2.1 prikazuje sestavo baze KEGG.

Category	Database	Content	Color
Systems information	KEGG PATHWAY	KEGG pathway maps	
	KEGG BRITE	BRITE functional hierarchies	
	KEGG MODULE	KEGG modules of functional units	
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups	
	KEGG GENOME	KEGG organisms with complete genomes	
	KEGG GENES	Gene catalogs of complete genomes	
	KEGG SSDB	Sequence similarity database for GENES	
Chemical information	KEGG COMPOUND	Metabolites and other small molecules	
	KEGG GLYCAN	Glycans	
	KEGG REACTION	Biochemical reactions	
	KEGG RPAIR	Reactant pair chemical transformations	
	KEGG RCLASS	Reaction class defined by RPAIR	
	KEGG ENZYME	Enzyme nomenclature	
Health information	KEGG DISEASE	Human diseases	
	KEGG DRUG	Drugs	
	KEGG DGROUP	Drug groups	
	KEGG ENVIRON	Crude drugs and health-related substances	

Slika 2.1: Baza KEGG. 17 manjših baz je razdeljenih na 4 skupine: sistemska, genomska, kemijska in zdravstvena skupina [21].

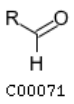
2.1.1 Uporabljeni podatki

Kot smo že omenili, osnovno podatkovno enoto pri našem delu predstavlja kemijska reakcija. Po definiciji v [33] je kemijska reakcija "proces, v katerem pride do trajne spremembe kemijskih in fizikalnih lastnosti snovi. V reakcijo vstopajoče snovi so reaktanti, izstopajoče snovi pa produkti. Kemijske reakcije zapisujemo s kemijskimi enačbami." Na podatkovni bazi KEGG je označena z nizom Rxxxxx, kjer črka R označuje, da gre za reakcijo, petmestna številka xxxxx pa predstavlja njeno unikatno številko. Slika 2.2 prikazuje kemijsko reakcijo primarnega alkohola.

Entry	R00623 Reaction
Name	primary_alcohol:NAD+ oxidoreductase
Definition	Primary alcohol + NAD+ <=> Aldehyde + NADH + H+
Equation	C00226 + C00003 <=> C00071 + C00004 + C00080
	

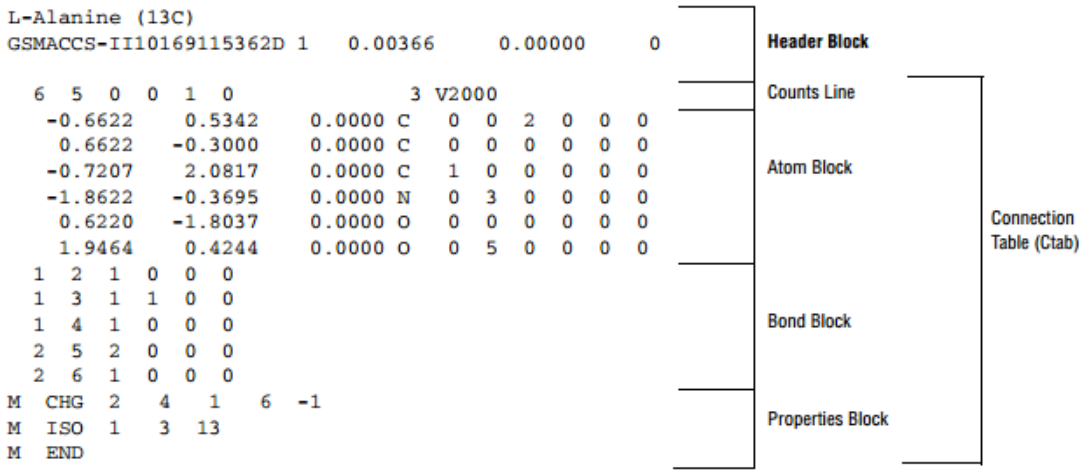
Slika 2.2: Zapis kemijske reakcije primarnega alkohola [21].

Za vsako kemijsko reakcijo potrebujemo njene reaktante in produkte, ki so sestavljeni iz molekul. Molekula je sestavljena iz dveh ali več atomov, ki jih povezujejo kemijske vezi. Na podatkovni bazi KEGG je molekula označena z nizom Cxxxxx, kjer črka C označuje, da gre za molekulo (angl. compound), petmestna številka xxxxx pa predstavlja njeno unikatno številko. Slika 2.3 prikazuje molekulo aldehida.

Entry	C00071 Compound
Name	Aldehyde; RCHO
Formula	CHOR
Structure	 C00071 Mol file KCF file DB search Jmol KegDraw
Comment	Generic compound in reaction hierarchy
Reaction	R00538 R00544 R00623 R00625 R00631 R00634 R00635 R00636 R00637 R00638 R00639 R01409 R01853 R01854 R02745 R03415 R03536 R06133 R07158 R07210 R07326 R07328 R08372 R09359 R09480 R10388 R10713 R10911 R10912
Pathway	map00071 Fatty acid degradation
Enzyme	1.1.1.1 1.1.1.2 1.1.1.71 1.1.2.7 1.1.3.13 1.1.9.1 1.1.99.20 1.2.1.3 1.2.1.4 1.2.1.5 1.2.3.1 1.2.5.2 1.2.7.5 1.2.99.6 1.2.99.7 1.4.3.4 1.4.3.5 1.4.3.8 1.4.3.10 1.4.3.12 1.4.3.21 1.4.3.22 1.7.3.1 1.14.14.3 1.14.14.5 3.3.2.2 3.3.2.5 4.1.1.1 4.1.2.10 4.1.2.11 4.1.2.47

Slika 2.3: Zapis molekule aldehida [21].

Vsako molekulo smo shranili lokalno na disk v molskem formatu datoteke. Molski format datoteke (MDL Molfile) je standardni format, ki hrani informacije o atomih, vezeh, povezanosti in koordinatah molekule. Slika 2.4 prikazuje molsko datoteko.



Slika 2.4: Zapis molske datoteke [21].

2.2 Tarčne skupine reakcij

V prvem delu diplomske naloge smo iskali najboljše parametre. To smo storili s pomočjo šestih osnovnih skupin kemijskih reakcij. Za končno analizo smo v drugem delu potrebovali skupino kemijskih reakcij proteinov, ki vstopajo v interakcijo z RNA.

2.2.1 Osnovne skupine kemijskih reakcij

Kemijske reakcije lahko razdelimo v različne skupine glede na več kriterijev: po izmenjavi snovi, reakcijskem mehanizmu, agregatnem stanju, vrsti prenešenih delcev, spremembi notranje energije, smeri [33]. V našem primeru smo izbrali kriterji po izmenjavi snovi, saj gre za najsplošnejši pristop

ločevanja kemijskih reakcij. Poleg tega so kemijske reakcije pravzaprav na vseh bioloških podatkovnih bazah razvrščene v skupine glede na kriterji izmenjave snovi, ki hkrati predstavljajo tudi šest osnovnih encimatskih skupin. Skupine se glede na različne kriterije med seboj bolj ali manj prekrivajo.

Razvrščanje kemijskih reakcij v 6 osnovnih skupin, po kriteriju izmenjave snovi, povzeto po [33]:

- **Razpad ali analiza** - iz enega reaktanta nastaneta dva ali več novih enostavnejših produktov
  Primer: $(NH_4)_2Cr_2O_7 + \Delta \rightarrow Cr_2O_3 + 4H_2O + N_2$
  To so kemijske reakcije, ki jih pospešujejo encimi iz oksidoreduktazne (angl. oxidoreductase) skupine (EC 1).

- **Enojna zamenjava ali substitucija** - enega od elementov v spojini zamenja drug, bolj reaktiven element.
  Primer: $CuSO_4 + Fe \rightarrow FeSO_4 + Cu$
  To so kemijske reakcije, ki jih pospešujejo encimi iz transferazne (angl. transferase) skupine (EC 2).

- **Izgorevanje** - molekula kisika se veže z drugo molekulo, tako da nastane molekula vode in molekula ogljikovega dioksida.
  Primer: $C_10H_8 + 12O_2 \rightarrow 10CO_2 + 4H_2O$
  To so kemijske reakcije, ki jih pospešujejo encimi iz hidrolazne (angl. hydrolase) skupine (EC 3).

- **Spajanje ali sinteza** - iz dveh ali več reaktantov nastane bolj kompleksen produkt.
  Primer: $N_2 + 3H_2 \rightarrow 2NH_3$
  To so kemijske reakcije, ki jih pospešujejo encimi iz liazne (angl. lyase) skupine (EC 4).

- **Dvojna zamenjava ali metateza** - dve snovi si izmenjata ione, tako da nastaneta dve novi snovi.
  Primer: $NaCl + AgNO_3 \rightarrow NaNO_3 + AgCl$
  To so kemijske reakcije, ki jih pospešujejo encimi iz izomerazne (angl. isomerase) skupine (EC 5).

- **Kislina-baza** - to je posebna vrsta dvojne zamenjave, ko med seboj zreagirata kislina in baza
  Primer: $HBr + NaOH \rightarrow NaBr + H_2O$
  To so kemijske reakcije, ki jih pospešujejo encimi iz ligazne (angl. ligase) skupine (EC 6).

2.2.2 Kemijske reakcije proteinov v interakciji z RNA

Iz Castello [14] smo pridobili seznam genov, za katere je bilo eksperimentalno dokazano, da so odgovorni za tvorjenje proteinov, ki se vežejo na eno- ali dvoverižen RNA (angl. RNA-binding proteins). Za pridobitev vseh kemijskih reakcij proteinov v interakciji z RNA smo za vsak protein iz podatkovne baze KEGG preverili ali vsebuje kakšen gen iz seznama Castello. Naredili smo seznam takšnih proteinov, ki so vsebovali po vsaj en gen iz seznama Castello. Pridobili smo 43 takih proteinov, za katere vemo, da opravljajo dvojno funkcijo. Še enkrat lahko omenimo, da sta ti dve funkciji naslednji: delovanje v metaboličnih kemijskih reakcijah kot encim in vstopanje v interakcijo z RNA. Ciljno skupino kemijskih reakcij proteinov v interakciji z RNA smo dobili tako, da smo na podatkovni bazi KEGG poiskali vse kemijske reakcije, v katerih nastopajo proteini iz prej pridobljenega seznama 43 proteinov.

2.3 Lastnosti pridobljenih podatkov

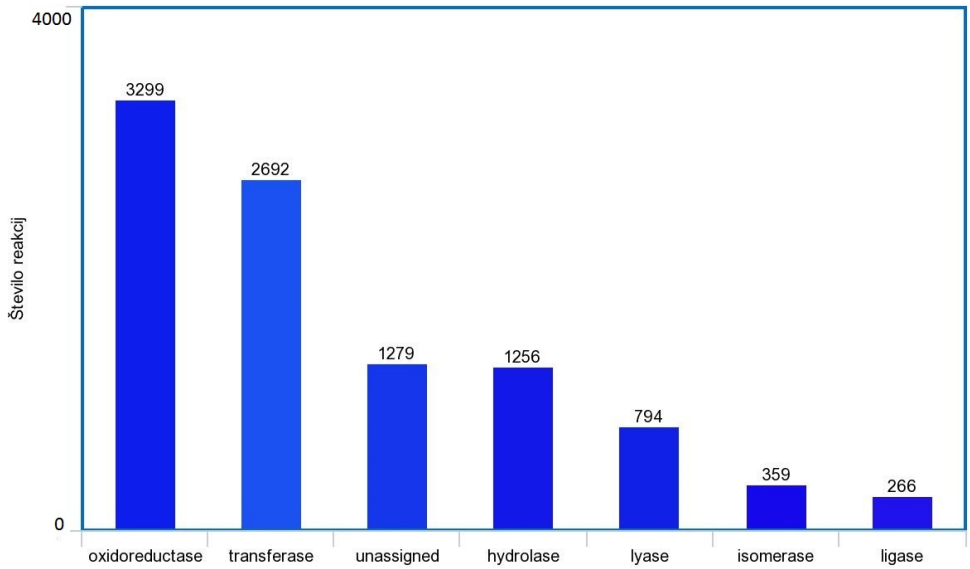
Tarčne skupine kemijskih reakcij smo razdelili v dve skupini (šest osnovnih skupin kemijskih reakcij in skupina kemijskih reakcij proteinov, ki vstopajo v interakcijo z RNA). Na tem mestu lahko povemo, da je slednja skupina, podmnožica prve skupine (vse reakcije, ki nastopajo v skupini kemijskih reakcij proteinov v interakciji z RNA prav tako nastopajo tudi v šestih osnovnih skupinah kemijskih reakcij).

- Število kemijskih reakcij osnovnih skupin

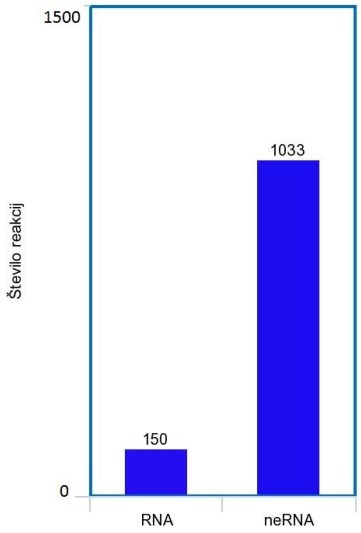
  Število vseh različnih kemijskih reakcij, na podatkovni bazi KEGG je 9945. Od te množice odstranimo 1279 reakcij, ki še nimajo določenega encimatskega razreda (angl. unassigned reactions). Odstranimo tudi tiste kemijske reakcije, ki vsebujejo molekulo s prazno molsko datoteko. Takšnih reakcij je 832. Na koncu nam za strojno učenje ostane 7834 kemijskih reakcij. Slika 2.5 prikazuje kemijske reakcije razvrščene v šest encimatskih skupin.

- Število kemijskih reakcij proteinov v interakciji z RNA

  Število takšnih kemijskih reakcij na podatkovni bazi KEGG je 150. Tiste kemijske reakcije, ki vsebujejo molekule s prazno molsko datoteko, zaradi pomanjkanja informacij iz množice odstranimo. Na koncu ostanejo 104 kemijske reakcije proteinov v interakciji z RNA. Slika 2.6 prikazuje dve skupini kemijskih reakcij, skupino kemijskih reakcij proteinov, ki vstopajo v interakcijo z RNA, in skupino kemijskih reakcij proteinov, ki ne vstopajo v interakcijo z RNA.



Slika 2.5: Število vseh kemijskih reakcij na podatkovni bazi KEGG. Razvrščene v šest encimatskih skupin, z dodatno skupino neuvrščenih kemijskih reakcij.



Slika 2.6: Število kemijskih reakcij proteinov v interakciji z RNA v primerjavi s kemijskimi reakcijami proteinov, ki ne vstopajo v interakcijo z RNA (homo sapiens).

Poglavje 3

Metode in orodja

Uporabili smo številne metode in algoritme iz kemoinformatike in strojnega učenja, da smo se uspešno približali zastavljenemu cilju. V pričujočem poglavju opišemo programska orodja in druge konstrukte, ki smo jih uporabili pri izdelavi končne napovedne metode.

3.1 Uporabljene knjižnice in orodja

Pri delu smo si pomagali z velikim številom odprtokodnih programskih orodij, knjižnic in paketov.

- **RDKit**
  RDKit je odprtokoden programski vmesnik, napisan v jezikih C++ in Python. Primeren je za izdelavo računalniških aplikacij v povezavi s kemijo in bioinformatiko. Največkrat se ga uporablja za obdelavo, manipulacijo in primerjavo kemijskih struktur. Poleg tega pa omogoča še veliko drugih funkcionalnosti:

  1. branje in pisanje kemijskih struktur v različne datotečne formate,
  2. razstavljanje molekul na fragmente,
  3. simulacijo kemijskih reakcij,

4. generiranje kemijskih profilov,
5. grafični prikaz molekul in drugih kemijskih struktur [23].

- **Scikit-learn**

  Je odprtokodna knjižica s podporo strojnega učenja, napisana v programskih jezikih Python in Cython. Ima vgrajene različne algoritme za klasifikacijo, regresijo in gručenje vključno z metodami, kot so metoda podpornih vektorjev, logistična regresija, naivni Bayesov klasifikator, naključni gozd, gradientni spust [20].

- **Unbalanced-dataset**

  Je odprtokodna knjižnica, napisana v programskem jeziku Python, ki ponuja veliko število metod za prevzorčenje neuravnoteženih podatkovnih množic. Prevzorčne tehnike so trenutno razdeljene v dve kategoriji, in sicer v tehnike nadvzorčenja in podvzorčenja [22].

- **Orange**

  Je odprtokodno orodje za strojno učenje in podatkovno rudarjenje. Je preprost in intuitiven program, saj omogoča pristop vizualnega programiranja. Primeren je za začetnike in zahtevnejše uporabnike [15]. Orodje Orange je napisano v programskih jezikih C++ in Python. Omogoča zelo veliko značilnih pristopov in metod, ki so nam znane s področja umetne inteligence, podatkovnega rudarjenja in strojnega učenja, kot so predprocesiranje podatkov, klasifikacija, regresija, gručenje podatkov, validacijske funkcije, ansambelske metode in drugo. Orodje je bilo razvito v Laboratoriju za bioinformatiko Fakultete za računalništvo in informatiko na Univerzi v Ljubljani [35].

3.2 Kemijski profili

Kemijski profil (angl. fingerprint) je abstraktna reprezentacija molekule, zapisana v binarnem bitnem nizu. Algoritem potrebuje za uspešno tvorjenje kemijskega profila (angl. fingerprinting algorithm) molekule naslednje informacije:

- vzorec za vsak atom,
- vzorec vsakega atoma in njegovih sosedov skupaj z vezjo, ki jih povezuje,
- vzorec skupin atomov, ki so med seboj povezani v razmaku dolžine dveh vezi,
- vzorec skupin atomov, ki so med seboj povezani v razmaku do največ sedem vezi [30].

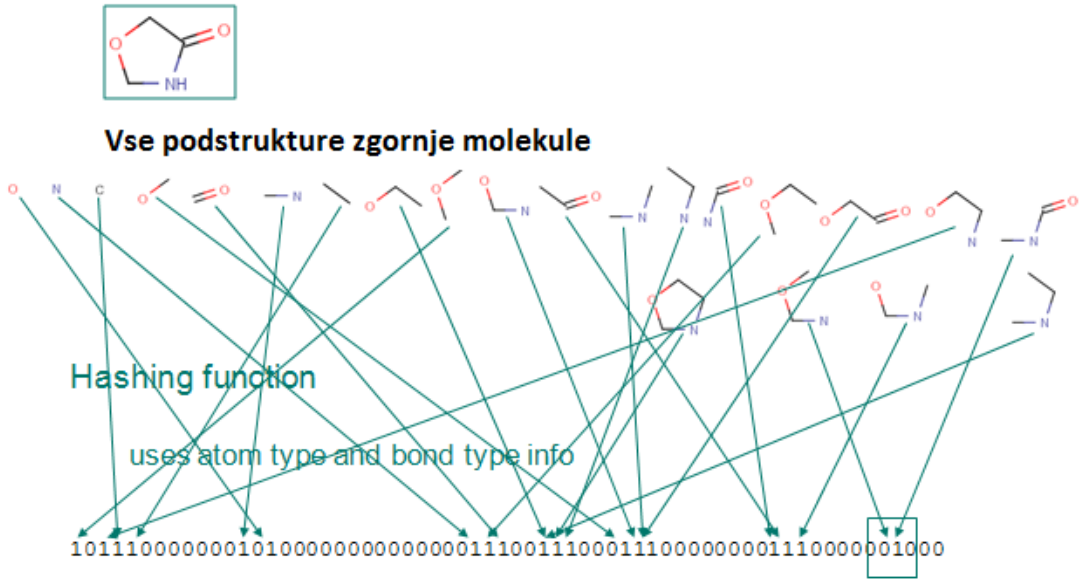
povezane vezi	**generirani vzorci**
0 povezanih vezi	O, N, C
1 povezana vez	OC, C=O, CN, CC
2 povezani vezi	CCO, COC, OCN, CC=O, CNC, CCN, NC=O
3 povezane vezi	COCC, OCC=O, OCCN, CNC=O, CNCC, CNCO, NCOC
4 povezane vezi	OCCNC

Tabela 3.1: Prikaz generiranih vzorcev za molekulo O=C1COCN1 [30].

Iz tabele 3.1 preštejemo, da se tvori 22 različnih vzorcev, s pomočjo katerih se nato tvori kemijski profil molekule O=C1COCN1. Algoritem pretvori posamezno molekulo v kemijski profil v štirih korakih:

1. Na podlagi parametra b (največji razmik med vezmi) detektira vzorce za vsako vrednost i = 0,...,b-1.

2. Za vsak vzorec detektira morebitne vejitvene točke (angl. branching points).
3. Detektira vse ciklične vzorce.
4. Z uporabo preslikovalne metode, postavi vnaprej določeno število bitov v kemijskem profilu (binarni bitni niz). Lahko se zgodi, da isti bit spreminja več vzorcev (angl. bit collision, razvidno iz slike 3.1) [31].

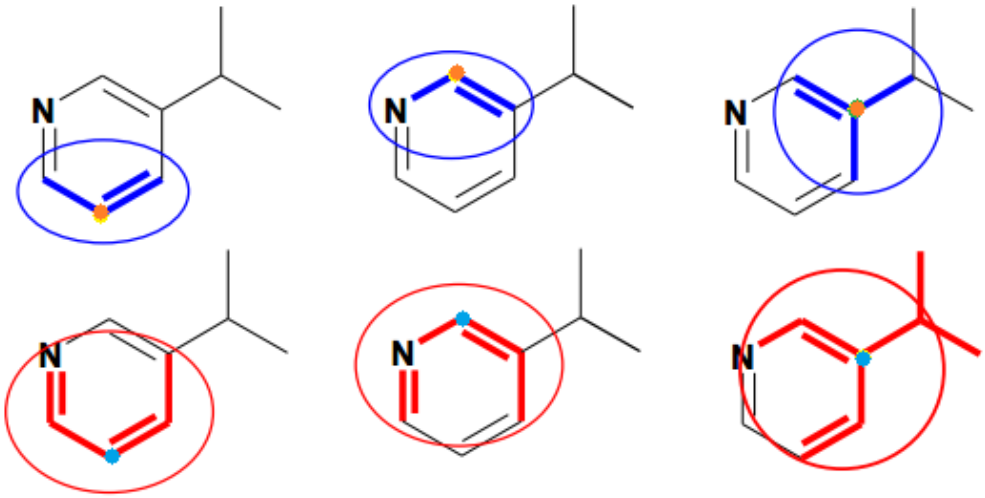


Slika 3.1: Prikaz generiranja kemijskega profila molekule O=C1COCN1. Slika povzeta po [31].

3.2.1 Vrste kemijskih profilov

Pri delu smo uporabili sedem načinov opisovanja kemijskih profilov. Od tega smo štiri kemijske profile generirali s kombiniranjem osnovnih treh. Osnovni trije kemijski profili temeljijo na prej opisanem postopku. Razlikujejo se predvsem v načinu generiranja posameznih vzorcev v molekuli. Pridobili smo jih iz knjižnice RDKit.

- Kemijski profil **Morgan**
  Pri Morganovem ali krožnem kemijskem profilu (angl. circular fingerprint) se molekularni vzorci generirajo glede na polmer okolice vsakega atoma v molekuli [2]. Na sliki 3.2 prikažemo generiranje vzorcev za Morganov kemijski profil.



Slika 3.2: Prikaz generiranja vzorcev glede na izbran polmer okrog izbranega atoma v molekuli. Modra barva (polmer = 1), rdeča barva (polmer = 2) [2].

- Kemijski profil **atomski par**
  Kemijski profil atomskih parov generira molekularne vzorce s kombiniranjem parov vseh atomov v molekuli skupaj z medsebojnimi vezmi in pi-elektroni.

- Kemijski profil **Morgan kot bitni vektor**
  Pri kemijskem profilu Morgan kot bitni vektor gre v osnovi za enako generiranje molekularnih vzorcev kot pri kemijskem profilu Morgan. Razlikujeta se v načinu shranjevanja bitov/števil v končnem kemijskem profilu. Medtem ko Morganov kemijski profil shranjuje v kemijski profil število pojavitev posameznih molekularnih vzorcev, se pri kemijskem profilu Morgan kot bitni vektor v kemijski profil shrani binarno vre-

dnost 0 ali 1, ki pove, ali je posamezen vzorec v kemijskem profilu prisoten ali ne.

3.2.2 Opisovanje kemijskih reakcij s kemijskimi profili

Opisali smo, kako tvorimo kemijski profil za posamezno molekulo posebej. Pri našem delu potrebujemo kemijski profil kemijske reakcije, ki je sestavljena iz večjega števila posameznih molekul. Kemijski profil posamezne kemijske reakcije pridobimo na dva načina.

- Opis kemijske reakcije z **razliko kemijskih profilov** produktov in reaktantov
  Če je kemijska reakcija stehiometrična (vsi atomi leve strani reakcije so prisotni tudi na desni strani reakcije), potem se bo razlika molekul reaktantov in molekul produktov kemijskih profilov odražala v spremembi kemijskih vezi, ki se spremenijo po poteku kemijske reakcije [30]. V našem primeru smo ravnali tako, da smo sešteli kemijske profile reaktantov in kemijske profile produktov ter naredili razliko vsot kemijskih profilov produktov in reaktantov, da smo dobili končno reprezentacijo kemijskega profila kemijske reakcije. Ta način smo lahko uporabili pri kemijskih profilih atomskih parov in Morgan, ki ga na sliki 3.3 predstavimo s programsko kodo.

- Opis reakcije z **logičnimi funkcijami kemijskih profilov** produktov in reaktantov
  Pri kemijskih profilih, ki so sestavljeni iz bitnih vektorjev (Morganov kemijski profil kot bitni vektor), tvorimo kemijski profil posamezne reakcije tako, da namesto seštevanja molekul reaktantov in seštevanja molekul produktov združujemo reaktante in produkte z logičnim operatorjem »OR«. Nato pa združimo reaktante in produkte v končni kemijski profil reakcije z logičnim operatorjem »XOR«. Razlog za tako predstavitev končnega kemijskega profila reakcije je v tem, da so kemij-

ski profili v tem primeru sestavljeni iz logičnih bitnih vektorjev in ne več iz vektorjev, ki nosijo celoštevilčne vrednosti. Če ne bi upoštevali tega pravila in bi še naprej seštevali in odštevali profile, bi lahko izgubili informacije, ki jih nosijo logični bitni vektorji.

```
for reactants, products in reactions:
  reactants_FP = [AllChem.GetMorganFingerprint \
  (Chem.MolFromMolFile('Mols/'+cid)) for cid in reactants]
  reac_FP_sum = None
  for fp in reactants_FP:
          if reac_FP_sum is None:
                  reac_FP_sum = fp
          else:
                  reac_FP_sum += fp

  products_FP = [AllChem.GetMorganFingerprint \
   (Chem.MolFromMolFile('Mols/'+cid)) for cid in products]
  prod_FP_sum = None

  for fp in products_FP:
        if prod_FP_sum is None:
                prod_FP_sum = fp
        else:
                prod_FP_sum += fp

  diff_FP = prod_FP_sum - reac_FP_sum
```

Slika 3.3: Programska koda za generiranje kemijskega profila kemijske reakcije z razliko kemijskih profilov produktov in reaktantov.

3.3 Modeliranje kemijskih reakcij s strojnim učenjem

Strojno učenje temelji na pridobivanju informacij iz podatkov. Pri našem delu uporabljamo standardne pristope in tehnike strojnega učenja, kot so predprocesiranje podatkov, uravnoteževanje učne množice, delitev podatkov na učno in testno množico, vrednotenje modelov.

3.3.1 Pregled napovednih modelov

Vse napovedne modele smo pridobili iz knjižnice scikit-learn.

- **Naključni gozd** (angl. random forest)
  Naključni gozd sodi med najbolj priljubljene napovedne modele, saj je preprost za uporabo (zahteva malo vhodnih parametrov), deluje dobro na velikih množicah podatkov in je hiter. Deluje tako, da zgradi n-dreves, pri čemer za vsako drevo uporabi približno 66% naključnih primerov iz celotne množice. Na vsakem vozlišču pri grajenju posameznega drevesa izbere naključno število atributov m iz celotnega nabora atributov, ki jih uporabi, da razdeli drevo naprej na nova vozlišča, kjer znova izbere m naključnih atributov. Na koncu vsako posamezno drevo klasificira testni primer v določen razred. Naključni gozd klasificira testni primer v tisti razred, ki je imel zbranih največ glasov med n-zgrajenimi drevesi. Naključni gozd se v primerjavi z drugimi napovednimi modeli izkaže za zelo uspešen algoritem.

- **Metoda podpornih vektorjev** (angl. support vector machine)
  Metoda podpornih vektorjev je klasifkator, ki poskuša ločiti med primeri dveh ali več razredov tako, da poskuša med njimi (primeri so predstavljeni kot točke na ravnini), narisati taki premici (podporna vektorja), da bo njuna medsebojna razdalja največja. Testni primer se preslika na isto ravnino, kjer se mu na podlagi razdalje do podpornih

vektorjev določi razred, ki mu pripada. Če na ravnini ne obstajata podporna vektorja, ki bi ustrezno ločevala razreda, poskuša algoritem preslikati primere v višjo dimenzijo (v našem primeru iz 2D-ravnine v 3D-prostor, v splošnem v mnogo višje dimenzije), kjer je zdaj z ravninama možno bolje ločiti med primeri obeh razredov. Ta preslikava se dela s pomočjo jedrnih funkcij (angl. kernel functions). Če je število atributov mnogo večje od števila primerov, bodo napovedi tipično slabše [27].

- **Naivni Bayes** (angl. naive Bayes)
  V strojnem učenju so naivni Bayesovi klasifikatorji družina preprostih verjetnostnih klasifikatorjev, ki temelijo na Bayesovem teoremu. Beseda naivni se uporablja zaradi tega, ker metoda predpostavlja, da so vsi atributi med seboj neodvisni. Kljub tej poenostavljeni predpostavki, se naivni Bayesov klasifikator odlično obnese pri klasificiranju dokumentov in filtriranju nezaželene elektronske pošte. V primerjavi z drugimi algoritmi strojnega učenja, je zelo hiter in potrebuje razmeroma malo učnih podatkov, da oceni pomembnost posameznih atributov [26].

- **Stohastični gradientni spust** (angl. stohastic gradient descent)
  V zadnjem desetletju velikost podatkovnih množic narašča hitreje kot pa hitrost procesorjev. Zmožnosti metod strojnega učenja danes omejuje predvsem računski čas in ne več problem premajhnih množic podatkov. Stohastični gradientni spust je optimizacijska metoda gradientnega spusta, ki poskuša minimizirati funkcijo izgube, ki je zapisana kot vsota odvedljivih funkcij [28]. Za razliko od gradientnega spusta stohastični gradientin spust ne potrebuje vseh podatkov učne množice v posamezni iteraciji, temveč naključno izbere samo po en primer iz učne množice, na podlagi katerega naredi posodobitev parametrov za minimizacijo funkcije izgube. Stohastični gradienti spust se je v zadnjem obdobju izkazal za učinkovit algoritem, saj se odlično skalira z

naraščanjem množic podatkov [4].

- **Odločitveno drevo** (angl. decision tree)
  Pogosto uporabljena metoda v podatkovnem rudarjenju in strojnem učenju je odločitveno drevo. Vse vrednosti atributov morajo biti predstavljene v diskretni obliki. Vsako vozlišče predstavlja po en atribut. Vsako podvozlišče predstavlja vrednost izhodne spremenljivke, ki razdeli primere na več podmnožic. Razdeljevanje drevesa na podmnožice se konča, ko ima podmnožica v posameznem vozlišču enake vrednosti ali ko razdeljevanje ne izboljša vrednosti napovedovanja. Ta proces (angl. top-down induction of decision trees) je primer požrešnega algoritma in je najznačilnejši način za učenje odločitvenih dreves iz podatkov [29].

3.3.2 Predprocesiranje podatkov

Pred začetkom procesiranja podatkov z metodami strojnega učenja je treba poskrbeti, da so podatki čim bolj dobro izbrani in pripravljeni. Najprej smo odstranili osamelce, nato pa smo poskrbeli za uravnoteževanje učne množice.

Iskanje osamelcev

Osamelec je učni primer, ki se zelo razlikuje od preostalih učnih primerov. Osamelci se pojavijo v podatkovnih množicah po navadi zaradi napak pri meritvah in zajemanju podatkov. Da se izognemo šumu na učnih primerih, je vedno dobro, da osamelce odstranimo iz množice podatkov. Osamelce smo odstranjevali s pomočjo orodja Orange, kjer smo uporabili metriko Z-vrednost (angl. Z-score). Z-vrednost lahko izračunamo po naslednji enačbi:

$$z = (X - \mu)/\sigma \tag{3.1}$$

kjer X predstavlja vrednost posameznega primera, μ povprečno vrednost, σ pa standardni odklon. Za vsak primer izračunamo evklidsko razdaljo do n-najbližjih sosedov. Vsakemu primeru glede na izračunano razdaljo

določimo Z-vrednost. Če je ta večja od naprej določene vrednosti z, tak primer označimo za osamelec. Za vnaprej določeno vrednost z smo izbrali vrednost 3,5, za n pa vrednost 3. Na celotni množici (7834 primerov kemijskih reakcij) smo identificirali 39 osamelcev in jih iz množice tudi odstranili.

Uravnoteževanje učne množice

Vse uravnoteževalne metode smo pridobili iz knjižnice unbalanced-datasets.

- **SMOTE** (angl. Synthetic Minority Over-sampling TEchnique)
  SMOTE je posebna vrsta nadvzorčenja (angl. oversampling) manjšinskega razreda, z generiranjem dodatnih primerov. Deluje tako, da se za vsak primer iz manjšinskega razreda določi k-najbližjih sosedov. Glede na potrebo nadvzorčenja manjšinskega razreda se naključno izbere določeno število sosedov od k-najbližjih sosedov za vsak primer iz manjšinskega razreda. Sintetični primer se generira z interpolacijo primerov manjšinskega razreda, tako da se vzame razliko atributov med primerom manjšinskega razreda in njegovim sosedom ter se jo pomnoži z naključno vrednostjo med 0 in 1 [16].

- **Near Miss**
  Z metodo Near-Miss poskušamo odstraniti primere večinskega razreda. Obstajajo 3 različice te metode: NearMiss-1, NearMiss-2 in NearMiss-3. Pri NearMiss-1 se odstrani primere večinskega razreda, ki so blizu nekaterim primerom manjšinskega razreda. Odstranijo se primeri, katerih povprečna razdalja do treh najbližjih primerov manjšinskega razreda je najkrajša. Pri NearMiss-2 se odstranijo primeri večinskega razreda, ki so blizu vsem primerom manjšinskega razreda. Ohranijo se primeri, katerih povprečna razdalja do treh primerov iz manjšinskega razreda je najdaljša. NearMiss-2 vedno podvzorči večinski razred do te mere, da vsebuje isto število primerov kot manjšinska množica. Pri NearMiss-3 se za vsak primer iz manjšinskega razreda poišče določeno

število najbližjih primerov iz večinskega razreda. Odstranijo se tisti, katerih razdalja do posameznih primerov manjšinskega razreda je najkrajša [3].

- **Povezave Tomek** (angl. Tomek links)
  Povezave Tomek lahko definiramo po naslednjem pravilu: imamo dva primera Ei in Ej, ki pripadata različnima razredoma, ter razdaljo med njima d(Ei,Ej). Par (Ei,Ej) je povezava Tomek, če ne obstaja nobeden tak primer El, da bi veljala neenakost d(Ei,El) $<$ d(Ei,Ej) ali d(Ej,El) $<$ d(Ei, Ej). Če Ei in Ej tvorita povezavo Tomek, potem vsaj eden od primerov predstavlja šum na podatkih ali pa sta oba primera robna primera. Povezavo Tomek lahko uprabimo kot metodo podvzorčenja ali pa kot metodo čiščenja podatkov. Če uporabimo povezavo Tomek kot metodo podvzorčenja, bodo odstranjeni samo primeri, ki pripadajo večinskemu razredu. V slednjem primeru, bosta odstranjena primera obeh razredov [16].

- **Neighborhood Cleaning Rule** (NCL)
  NCL uporablja popravljeno Wilsonovo pravilo najbližjih sosedov (Wilson's Edited Nearest Neighbor Rule (ENN)) za odstranjevanje primerov večinsekga razreda. ENN odstrani vsak primer, če se razlikuje v najmanj dveh primerih od treh najbližjih sosedov po oznaki razreda. Za primer klasificiranja primerov dveh razredov, opišemo algoritem NCL po naslednjem pravilu: za vsak primer Ei iz učne množice, poiščemo tri najbližje sosede. Če Ei pripada večinskemu razredu, njegovi sosedi pa klasificirajo primer Ei kot primer iz manjšinskega razreda, potem Ei odstranimo iz učne množice. Če Ei pripada manjšinskemu razredu, njegovi sosedi pa klasificirajo primer Ei kot primer iz večinskega razreda, odstranimo njegove sosede [16].

- **SMOTE + povezave Tomek**
  Predstavljamo si množico podatkov, ki je sestavljena iz primerov dveh razredov, nekaj primerov manjšinskega razreda pa je globoko vsebovanih v gruči primerov večinskega razreda. Pri nadvzorčenju manjšinskega razreda z metodo SMOTE se bo nedoločenost množice še povečala, saj se bodo na novo generirali tudi »slabi« primeri, ki so pomešani med primeri večinskega razreda. Podobno velja tudi obratno, če so primeri večinskega razreda pomešani med primeri manjšinsekga razreda. Ideja, ki stoji v združevanju obeh metod, je ta, da nad učno množico, nadvzorčeno s SMOTE, poženemo metodo povezave Tomek, ki v tem primeru služi kot metoda za čiščenje podatkov. Namesto da odstranimo samo primere večinskega razreda, tedaj odstranimo tudi »slabe« primere obeh razredov. Združeni metodi SMOTE + povezave Tomek, so bili prvič uporabljeni prav na področju bioinformatike za izboljšavo napovedovanja številk EC proteinom [16].

- **SMOTE + ENN**
  Ideja združevanja metod SMOTE + ENN je podobna kot zgoraj opisana, le da metoda ENN odstrani večje število slabih primerov kot metoda povezav Tomek [16].

- **Naključno podvzorčenje** (angl. random undersapling)
  Naključno podvzorčenje je metoda brez uporabe hevristike, ki uravnoteži množici razredov tako, da naključno odstrani določeno število primerov iz večinske množice.

- **Naključno nadvzorčenje** (angl. random oversampling)
  Naključno podvzorčenje je metoda brez uporabe hevristike, ki uravnoteži množici razredov tako, da naključno replicira določeno število primerov iz manjšinskega razreda.

3.3.3 Ocenjevanje atributov

Izbira atributov (angl. feature selection) je proces tvorjenja podmnožice relevantnih atributov. Podatki pogosto vsebujejo veliko število atributov, ki so redundantni ali irelevantni in jih zaradi tega odstranimo iz množice atributov brez prevelike izgube informacij. Z odstranjevanjem nepomembnih atributov zmanjšamo čas, ki ga napovedni modeli potrebujejo za učenje, in prilagajanje napovednih modelov na nepomembne podatke (zmanjšamo varianco) [32]. V našem primeru smo uporabili metodo izbire atributov zato, da smo pridobili seznam najpomembnejših atributov. Torej seznam tistih atributov (vzorcev ali substruktur), ki so skupni največjemu številu primerov kemijskih reakcij proteinov v interakciji z RNA. Uporabili smo naključni gozd, ki ga lahko poleg metode za strojno učenje hkrati uporabimo tudi za izbiro pomembnih atributov. Metoda je dostopna v knjižnici scikit-learn.

- **Povprečno zmanjšanje nečistosti** (angl. mean decrease impurity)
  Naključni gozd je sestavljen iz večjega števila odločitvenih dreves. Vsako vozlišče v odločitvenem drevesu predstavlja pogoj posameznega atributa, ki poskuša razdeliti učne primere tako, da primere istega razreda klasificira v iste podskupine. Mera na podlagi katere je lokalno izbran optimalni pogoj, se imenuje nečistost. Za napoved se po navadi uporabi Ginijev indeks ali entropija, pri regresiji pa varianca. Pri učenju drevesa se izračuna, za koliko posamezen atribut zmanjša nečistost drevesa. Za vsak atribut pridobimo vrednost povečanja nečistosti drevesa. Na podlagi te mere določimo pomembnost vsakega posameznega atributa [25].

- **Povprečno zmanjšanje točnosti** (angl. mean decrease accuracy)
  Uporabimo lahko tudi metodo povprečnega zmanjšanja točnosti, kjer se permutirajo vrednosti posameznega atributa, nato pa se izračuna za koliko permutacije zmanjšajo točnost modela. Če je atribut nepomemben, se točnost napovedi ne bo spremenila veliko. Bolj kot je

atribut pomemben, večje bo zmanjšanje točnosti, potem ko ga permutiramo [25].

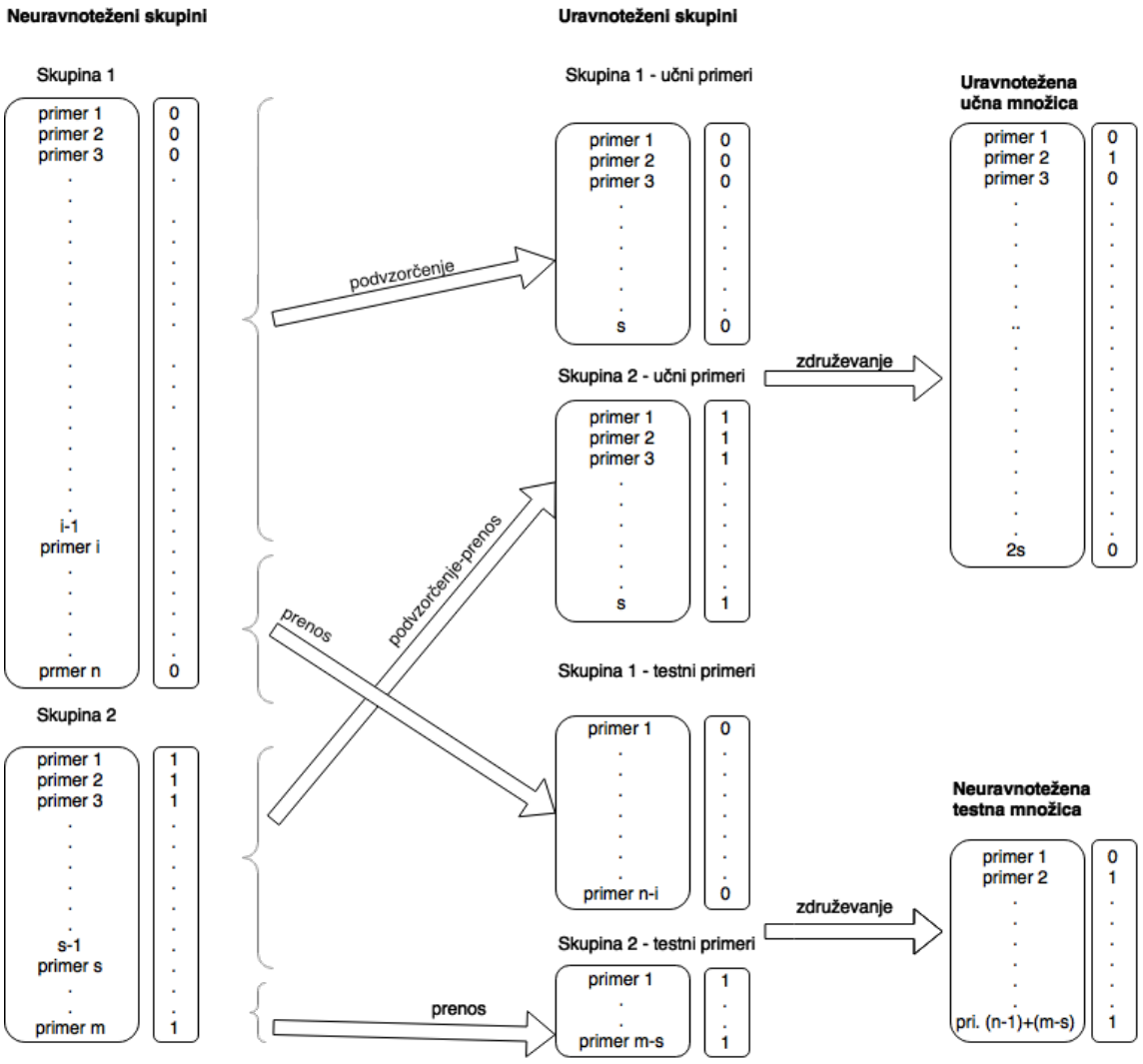
3.3.4 Delitev množice podatkov na učno in testno množico

V našem primeru smo učno in testno množico gradili iz množice podatkov, ki je vedno vsebovala dva razreda.

- **Prečno preverjanje** (angl. cross validation)
  Pri prečnem preverjanju razdelimo množico primerov na k-enakih delov. Posamezni deli se med seboj vzajemno izključujejo. V vsaki ponovitvi izberemo k-1 delov za treniranje modela in preostali del za testiranje modela. Na koncu je vsak del za testiranje izbran natanko enkrat. Rezultate napovedovanja na testni množici povprečimo, da dobimo končno izračunano natančnost napovednega modela.

- **Stratificirano naključno vzorčenje** (angl. stratified random sampling)
  Pri statificiranem naključnem vzorčenju množico vseh primerov razdelimo na ločene skupine po oznaki razreda. Iz vsake skupine naključno izberemo določeno število primerov v enakih proporcih kot so ti razporejeni v celotni množici. S tem ohranimo enako razmerje primerov enega in drugega razreda v učni in testni množici.

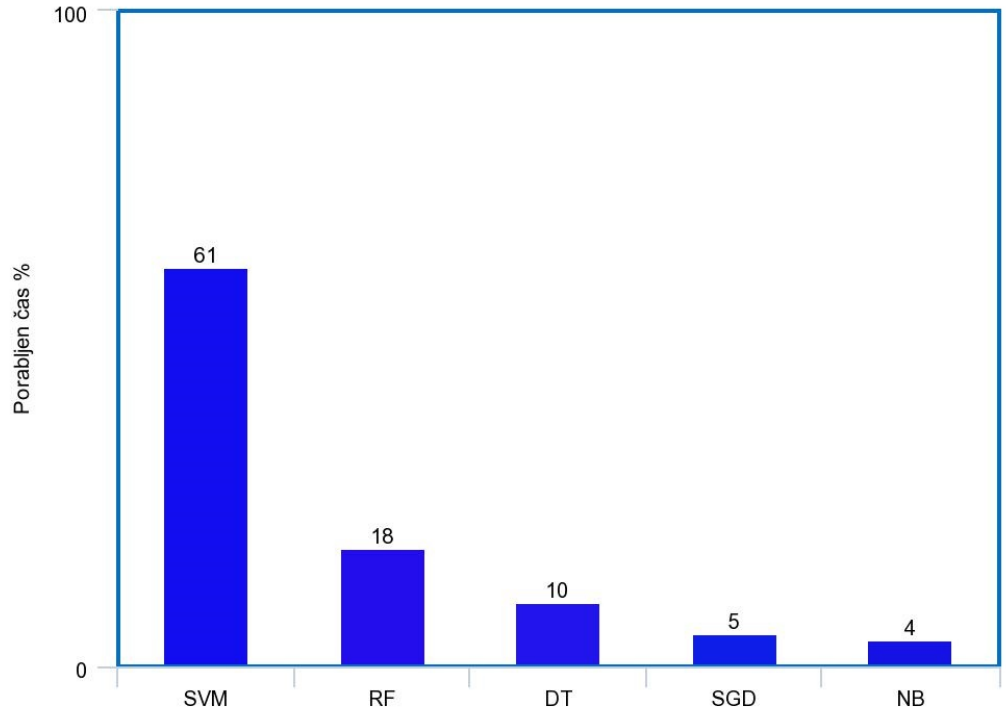
Slika 3.4 prikazuje shemo razdeljevanja podatkov na učno in testno množico, ki smo jo uporabili za določanje uravnotežene učne množice in neuravnotežene testne množice.



Slika 3.4: Prikaz združevanja dveh skupin kemijskih reakcij na testno in učno množico.

3.3.5 Primerjava modelov

Napovedne modele smo med seboj primerjali z uporabo metode, ki temelji na Friedmanovem testu in je opisana v [17]. V osnovi deluje tako, da vsakemu napovednemu modelu določi število, ki je enako osvojenemu mestu, glede na napovedno vrednost AUC med vsemi napovednimi modeli. Testiranje se izvede na različnih množicah podatkov. Uvrstitve posameznih modelov se povprečijo. Model z najnižjo povprečno vrednostjo posameznih uvrstitev je najboljši, model z najvišjo pa najslabši.



Slika 3.5: Porabljen čas pri strojnem učenju za učenje in napovedovanje.

Poglavje 4

Rezultati

V pričujočem poglavju predstavimo rezultate našega dela. Ponovno lahko povemo, da smo v prvem delu s pomočjo osnovnih skupin kemijskih reakcij iskali najboljši nabor parametrov, ki smo ga v drugem delu uporabili za končno analizo skupine kemijskih reakcij proteinov v interakciji z RNA.

4.1 Izbira najučinkovitejše uravnoteževalne metode

Pri izbiri najučinkovitejše uravnoteževalne metode smo med seboj primerjali sedem različnih metod, ki so dostopne v knjižnici Unbalanced-dataset. Metode, ki smo jih med seboj primerjali so: naključno podvzorčenje, NCL, povezave Tomek, SMOTE, SMOTE+povezave Tomek, SMOTE+ENN in Near miss. Testirali smo jih na dveh najmanj uravnoteženih in dveh najbolj uravnoteženih skupinah osnovnih kemijskih reakcij. Rezultati iz tabele 4.1 so pokazali, da se uravnoteževalne metode med seboj z izjemo metode Near miss, ne razlikujejo bistveno. Za nadaljne napovedne poskuse izberemo metodo naključnega podvzorčenja.

Razloga za odločitev sta:

1. množici kemijskih reakcij proteinov v interakciji z RNA sta zelo neuravnoteženi (rezultati naključnega podvzorčenja so najboljši ravno na najmanj uravnoteženih množicah),

2. naključno podvzorčenje je v primerjavi z ostalimi metodami, mnogo hitrejša metoda.

Uravnoteževalne metode	**Neuravnoteženi skupini**	**Uravnoteženi skupini**
Naključno podvzorčenje	**0,911**	0,910
CNN	0,906	**0,923**
Povezave Tomek	0,903	0,981
SMOTE+povezave Tomek	0,895	0,916
SMOTE	0,903	0,922
SMOTE+ENN	0,857	0,917
Near Miss	0,730	0,879

Tabela 4.1: Prikaz povprečnih vrednosti AUC pri izbiri različne uravnoteževalne metode. Testiramo jih na dveh najmanj uravnoteženih skupinah (ligaze 266 primerov in oksidoreduktaze 3299 primerov) in dveh najbolj uravnoteženih skupinah (ligaze 266 primerov in izomeraze 359 primerov). Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1.

4.2 Izbira najučinkovitejšega kemijskega profila in napovednega modela

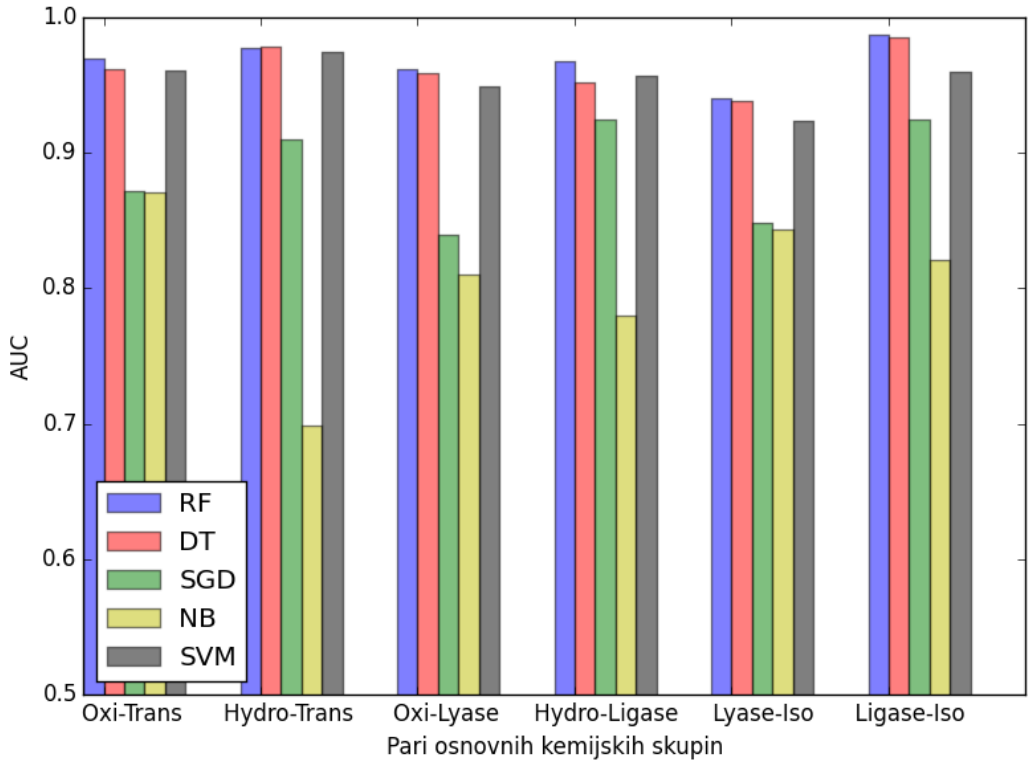
Pri izbiri najučinkovitejšega kemijskega profila smo med seboj primerjali sedem kemijskih profilov. Tri posamezne kemijske profile predstavlja množica: Morgan, Morgan kot bitni vektor in atomski par. Preostale štiri kemijske profile pa smo dobili s kombiniranjem posameznih kemijskih profilov: Morgan+MorganBitniVektor, Morgan+AtomskiPar, MorganBitniVektor+AtomskiPar, Morgan+MorganBitniVektor+AtomskiPar. Kemijske profile smo testirali na podatkih osnovnih skupin kemijskih reakcij. Vsaka kemijska skupina nastopa v dveh od šestih parov skupin. Pri vsakem testiranju kemijskega profila smo hkrati preverjali uspešnost napovednih modelov. Uporabili smo diagrame kritičnih razlik (CD, angl. Critical Difference), ki temelijo na Friedmanovem testu in so dostopni v paketu Orange. Iz diagramov kritičnih razlik (Slike 4.2, 4.4, 4.6, 4.8, 4.10, 4.12, 4.14) vidimo, da naključni gozd v šestih primerih zasede prvo mesto. V enem primeru pa si deli prvo mesto z metodo podpornih vektorjev. Za najučinkovitejši napovedni model tako izberemo metodo naključni gozd.

Na slikah 4.1, 4.3, 4.5, 4.7, 4.9, 4.11, 4.13 vidimo primerjavo uspešnosti kemijskih profilov. Kljub minimalnim razlikam med kemijskimi profili za najučikovitejši kemijski profil izberemo kombinacijo Morgan+MorganBitniVektor. V tabeli 4.2 prikažemo primerjavo kemijskih profilov.

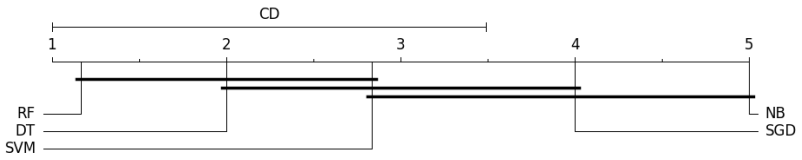
Kemijski profil	**AUC**
Morgan	0,967
MorganBitniVektor	0,971
AtomskiPar	0,955
Morgan+MorganBitniVektor	**0,972**
MorganBitniVektor+AtomskiPar	0,960
Morgan+AtomskiPar	0,965
Morgan+MorganBitniVektor+AtomskiPar	0,970

Tabela 4.2: Prikaz povprečnih vrednosti AUC sedmih kemijskih profilov. Za napovedno metodo uporabimo naključni gozd. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje. Povprečne vrednosti smo izračunali na podlagi šestih parov osnovnih kemijskih skupin.

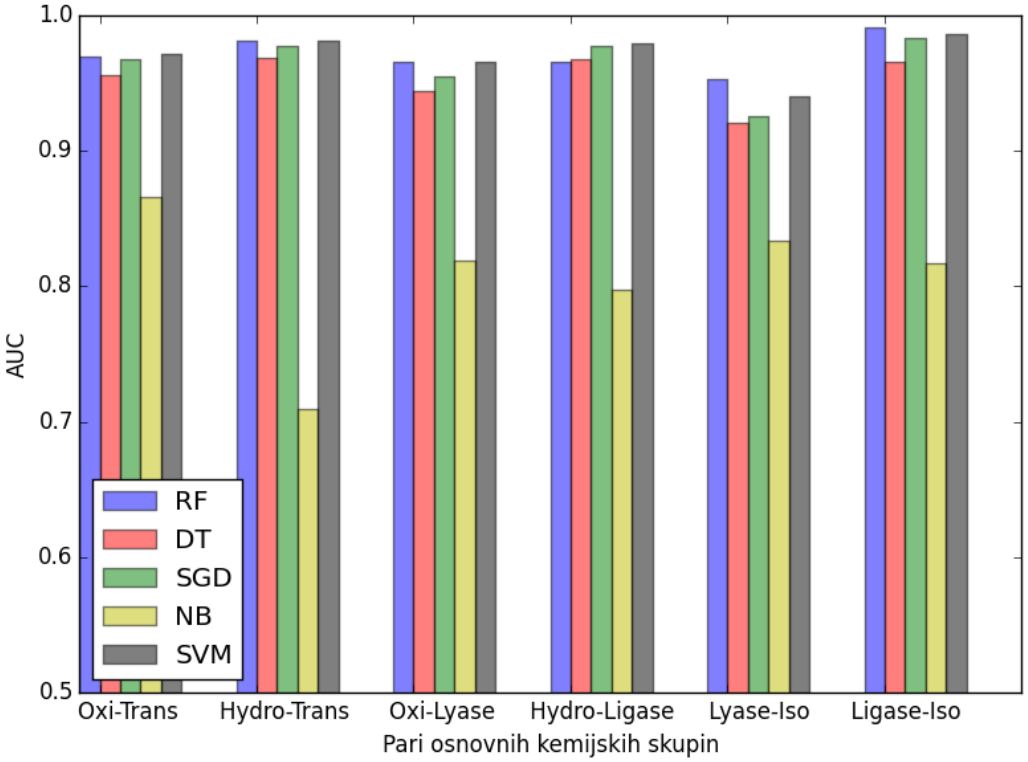
4.2.1 Posamezni kemijski profili



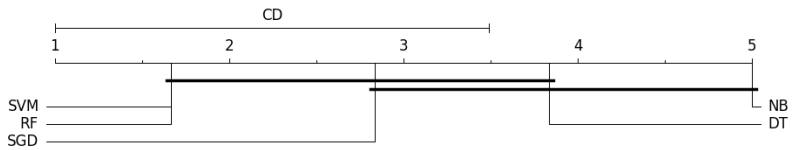
Slika 4.1: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kemijskega profila Morgan. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.



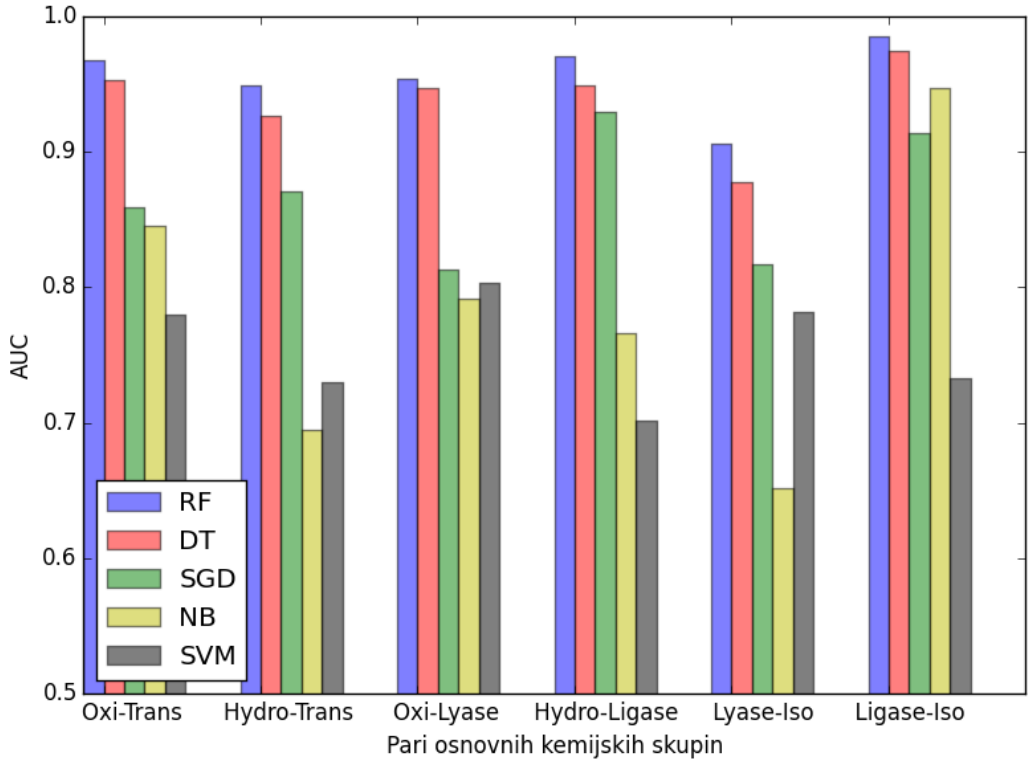
Slika 4.2: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kemijski profil Morgan.



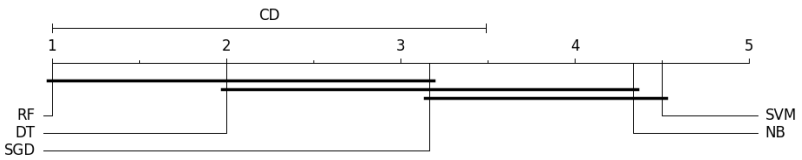
Slika 4.3: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kemijskega profila Morgan kot bitni vektor. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.



Slika 4.4: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kemijski profil Morgan kot bitni vektor.

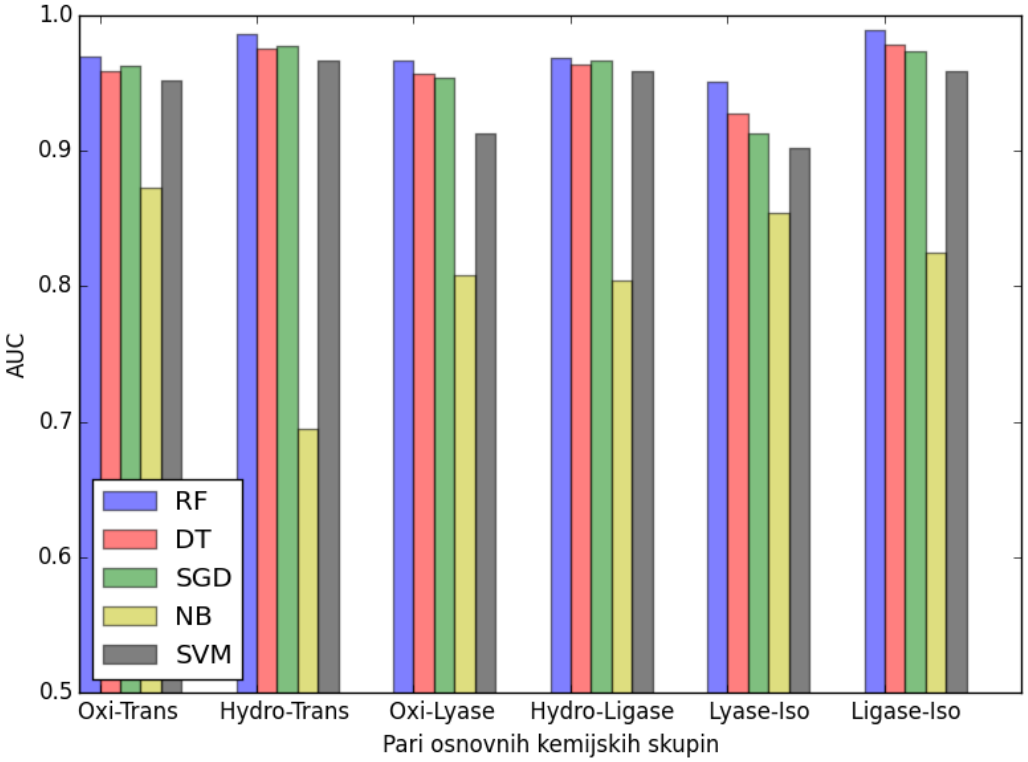


Slika 4.5: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kemijskega profila atomskih parov. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.

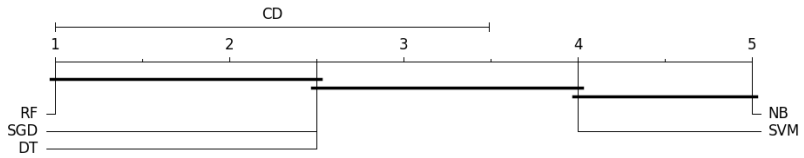


Slika 4.6: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kemijski profil atomskih parov.

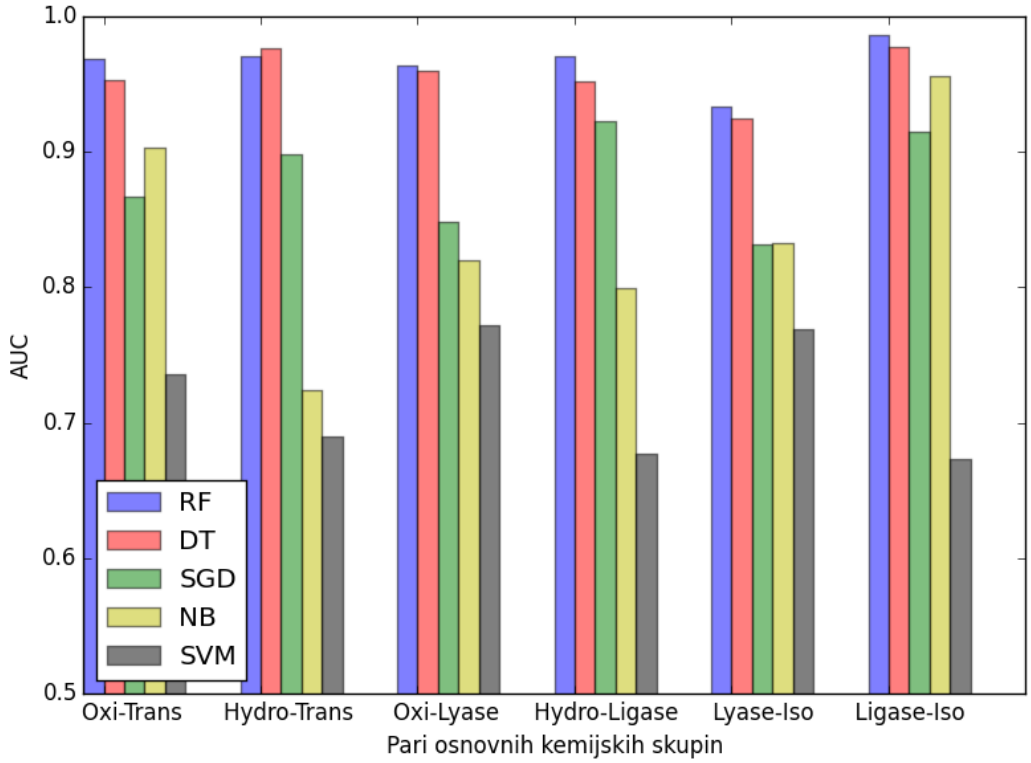
4.2.2 Kombinirani kemijski profili



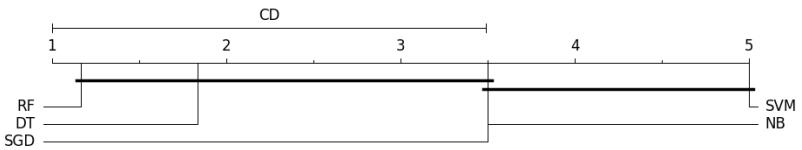
Slika 4.7: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kombinacije kemijskih profilov Morgan+MorganBitniVektor. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.



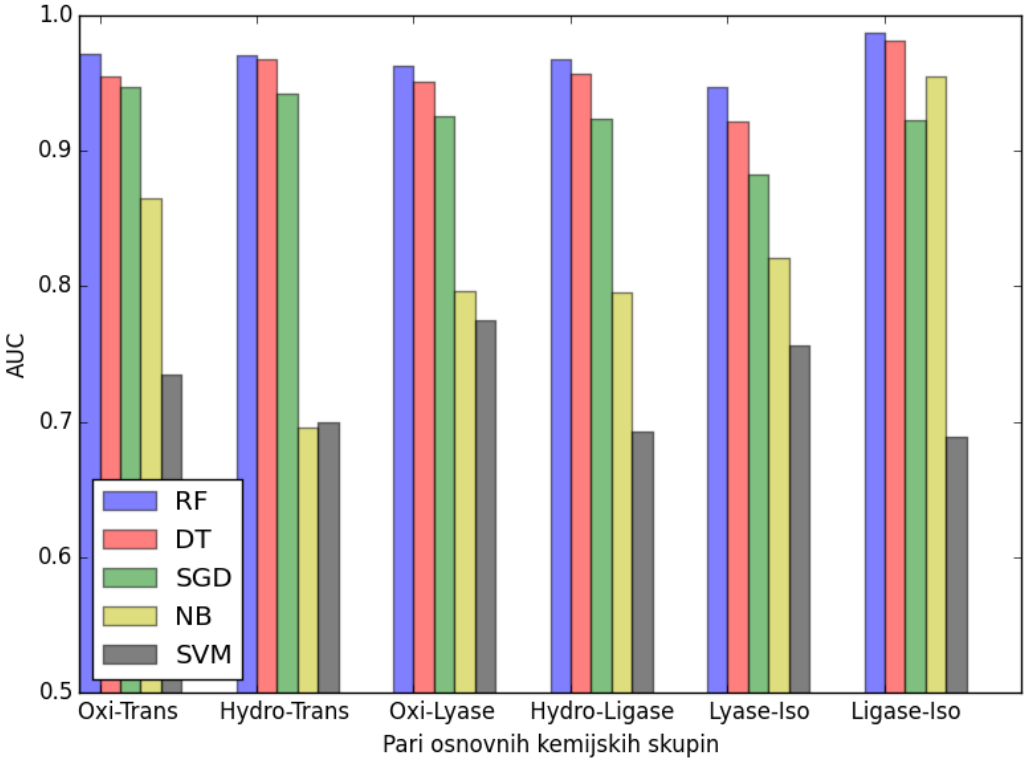
Slika 4.8: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kombiniran kemijski profil Morgan+MorganBitniVektor.



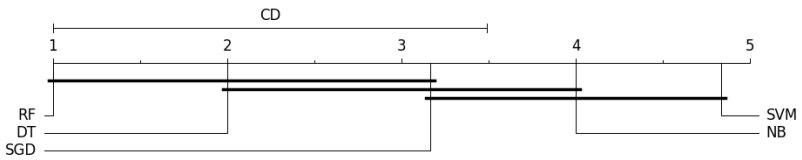
Slika 4.9: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kombinacije kemijskih profilov Morgan+AtomskiPar. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.



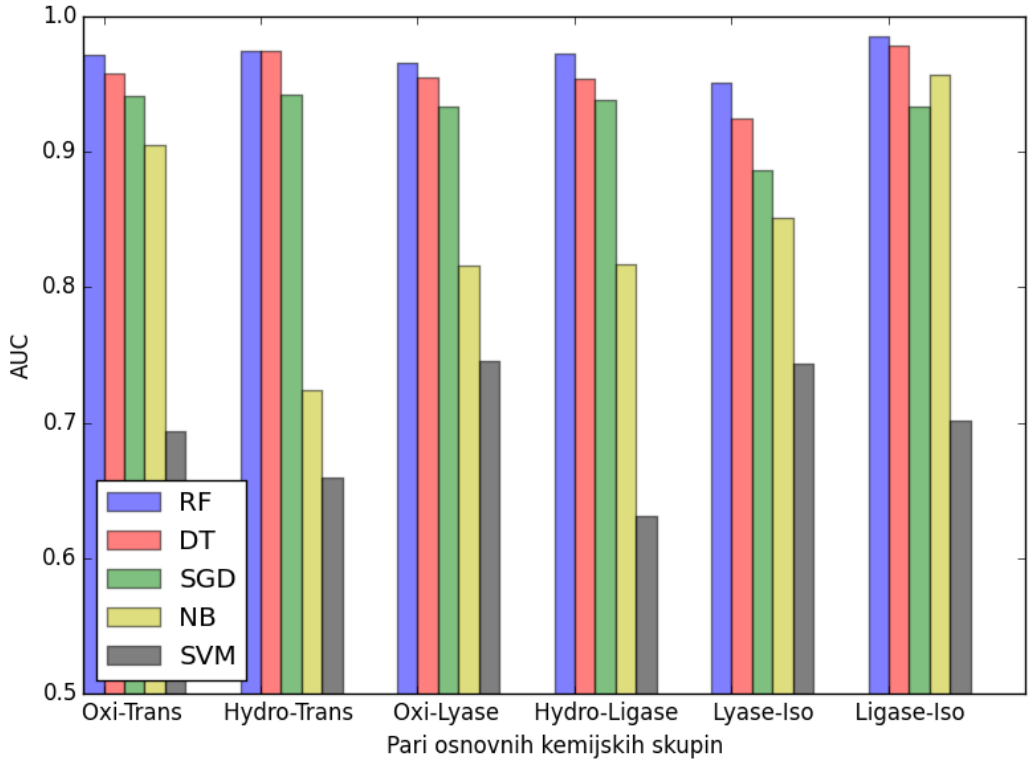
Slika 4.10: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kombiniran kemijski profil Morgan+AtomskiPar.



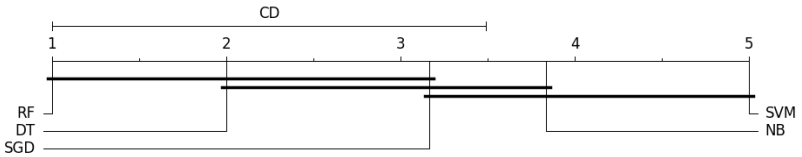
Slika 4.11: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kombinacije kemijskih profilov MorganBitniVektor+AtomskiPar. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.



Slika 4.12: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kombiniran kemijski profil MorganBitniVektor+AtomskiPar.



Slika 4.13: Prikaz vrednosti AUC, šestih parov osnovnih skupin kemijskih reakcij, z uporabo kombinacije kemijskih profilov Morgan+MorganBitniVektor+AtomskiPar. Testiramo 5 napovednih modelov. Uporabimo proporcijski test (30 iteracij) z razdeljevanjem učne in testne množice v razmerju 4:1. Za uravnoteževalno metodo učne množice izberemo naključno podvzorčenje.



Slika 4.14: Primerjava uspešnosti petih napovednih modelov s prikazom diagrama kritičnih razlik za kombiniran kemijski profil Morgan+MorganBitniVektor+AtomskiPar.

4.3 Napovedna uspešnost strojnega učenja kemijskih reakcij proteinov v interakciji z RNA

Empirično dokazano najboljše vhodne parametre, sedaj uporabimo pri klasificiranju kemijskih reakcij proteinov v interakciji z RNA. Rezultat končnega klasificiranja prikažemo v tabeli 4.3.

Kemijski profil	**Napovedna metoda**	**Uravnoteževalna metoda**	**AUC**
Morgan+MorganBitVektor	Naključni gozd	Naključno podvzorčenje	0,768

Tabela 4.3: Končni rezultat strojnega učenja kemijskih reakcij proteinov v interakciji z RNA. Uporabimo proporcijski test (50 iteracij), razmerje učne in testne množice 4:1. Kemijski profil Morgan+MorganBitniVektor, napovedna metoda naključni gozd ter uravnoteževalna metoda naključnega podvzorčenja

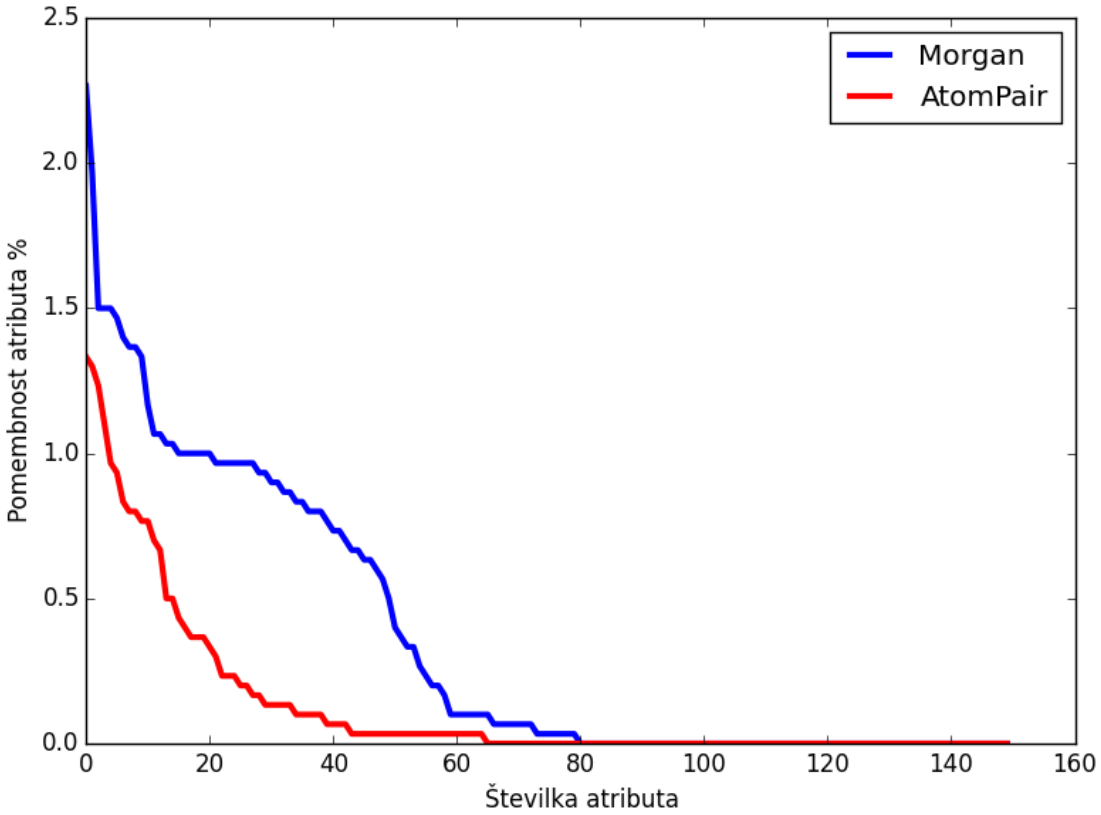
Za primerjavo uspešnosti rezultatov iz tabele 4.3 smo izvedli dodatno testiranje. Generirali smo umetni množici podatkov. Sintetična skupina 1 vsebuje 104 primere kemijskih reakcij. Sintetična skupina 2 vsebuje 1033 primerov kemijskih reakcij (enako kot RNA in ne-RNA, prikazano na sliki 2.6). Poskrbeti smo morali tudi zato, da obe sintetični skupini vsebujeta enako število kemijskih reakcij iz osnovnih skupin (enako kot RNA in ne-RNA). Edina razlika je bila ta, da smo sintetični množici izbirali naključno med vsemi primeri kemijskih reakcij. V tabeli 4.4 prikažemo rezultate dodatnega testiranja.

Kemijski profil	**Napovedna metoda**	**Uravnoteževalna metoda**	**AUC**
Morgan+MorganBitVektor	Naključni gozd	Naključno podvzorčenje	0,612

Tabela 4.4: Dodatno testiranje na naključno izbrani množici reakcij. Uporabimo proporcijski test (50 iteracij), razmerje učne in testne množice 4:1. Kemijski profil Morgan+MorganBitniVektor, napovedna metoda naključni gozd ter uravnoteževalna metoda naključnega podvzorčenja.

4.4 Pomembne strukturne spremembe v kemijskih reakcijah

Kot smo zapisali v poglavju 3.3.3 smo pomembne strukturne spremembe kemijskih reakcij pridobili z uporabo metode naključnega gozda, ki jo v tem primeru uporabimo za iskanje pomembnih atributov (podstruktur, vzorcev). Nad učno množico, ki je sestavljena iz kemijskih reakcij proteinov v interakciji z RNA in kemijskih reakcij proteinov, ki niso v interakciji Z RNA, poženemo omenjeno metodo. Metoda (feature_importances) iz knjižnice scikit-learn vrne seznam vseh atributov s pripadajočo pomembnostjo. Iz slike 4.15 vidimo, da so atributi med seboj različno pomembni.

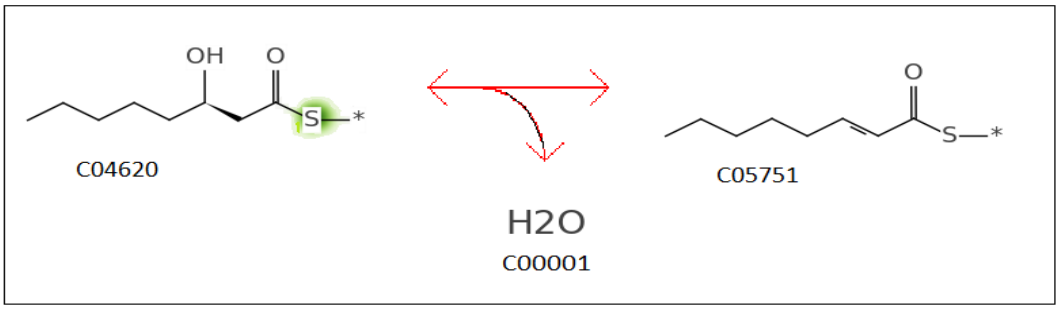


Slika 4.15: Pomembnost atributov za kemijska profila Morgan in atomski par. Uporabili smo metodo naključni gozd (povprečno zmanjšanje nečistosti). Atribute smo uredili po padajočih vrednostih.

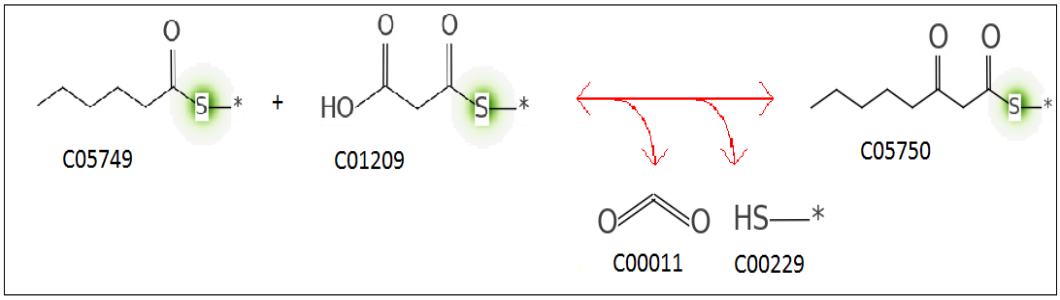
Strukturna sprememba	**Pomembnost [%]**
SC(C)=O	2,26
SC(=O)CC	1,96
OH	1,59
CC(=O)CC(=O)S	1,5
CCC(=O)CC	1,46
SC(=O)C=C	1,4
CCC(=O)SC	1,35
CC(C)=O	1,33
CSC(C)=O	1,06
CCCC(C)=O	1,06
CCC	1,03
CCSC(C)=O	1,03

Tabela 4.5: Najpomembnejše strukturne spremembe (atributi), ki nastajajo pri kemijskih reakcijah proteinov v interakciji z RNA.

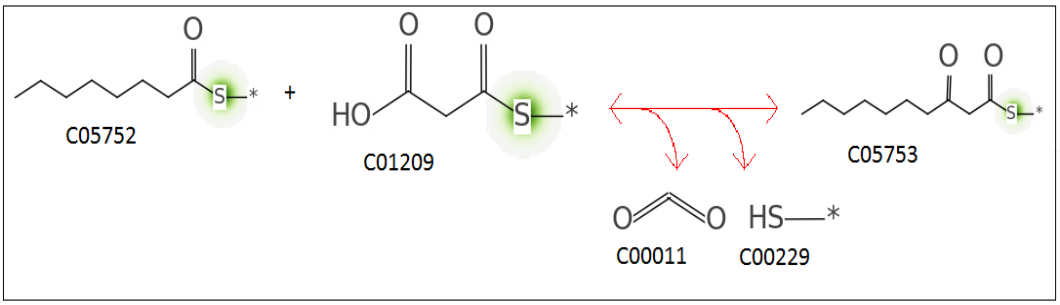
Na naslednjih slikah prikažemo tri najpomembnejše strukturne spremembe (SC(C)=O, SC(=O)CC, OH), ki nastanejo pri kemijskih reakcijah proteinov v interakciji z RNA. Pridobimo jih iz tabele 4.5. Posamezno strukturno spremembo prikažemo v treh različnih kemijskih reakcijah (prisotne so v več kot treh kemijskih reakcijah). Slike 4.16, 4.17, 4.18 prikazujejo kemijske reakcije, v katerih nastopa najbolj pomembna strukturna sprememba (SC(C)=O). Slike 4.19, 4.20, 4.21 prikazujejo kemijske reakcije, v katerih nastopa druga najbolj pomembna strukturna sprememba (SC(=O)CC). Slike 4.22, 4.23, 4.24 prikazujejo kemijske reakcije, v katerih nastopa tretja najbolj pomembna strukturna sprememba (OH).



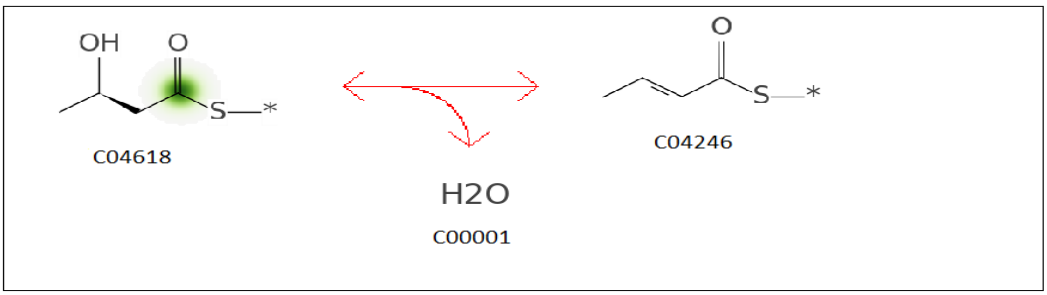
Slika 4.16: Prikaz kemijske reakcije R04537: C04620 <= >C05751 + C00001. Strukturna sprememba je v reakciji obarvana zeleno (SC(C)=O).



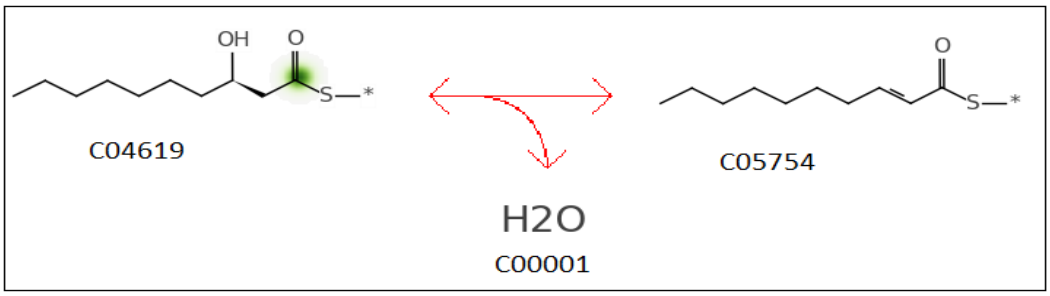
Slika 4.17: Prikaz kemijske reakcije R04957: C05749 + C01209 <= >C05750 + C00011 + C00229. Strukturna sprememba je v reakciji obarvana zeleno (SC(C)=O).



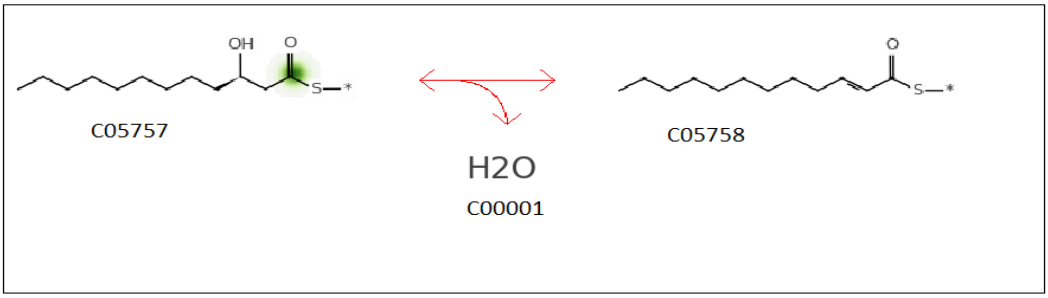
Slika 4.18: Prikaz kemijske reakcije R04960: C05752 + C01209 <= >C05753 + C00011 + C00229. Strukturna sprememba je v reakciji obarvana zeleno (SC(C)=O).



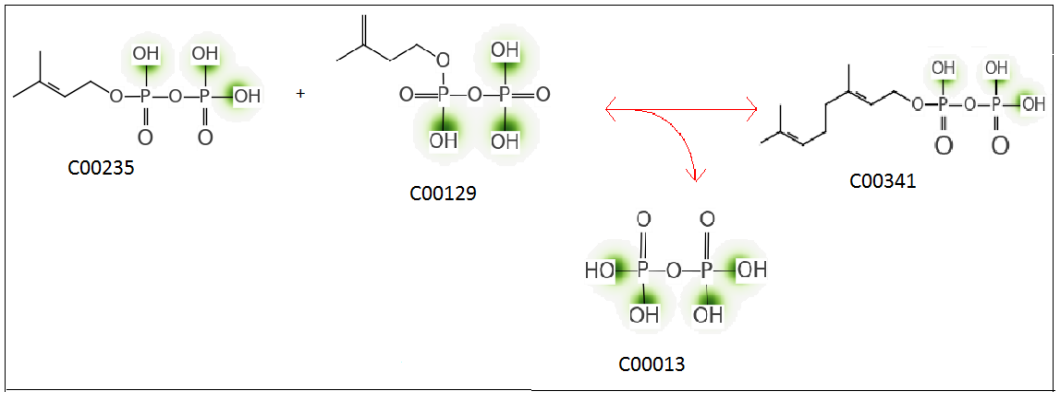
Slika 4.19: Prikaz kemijske reakcije R04428: C04618 <= >C04246 + C00001. Strukturna sprememba je v reakciji obarvana zeleno (SC(=O)CC).



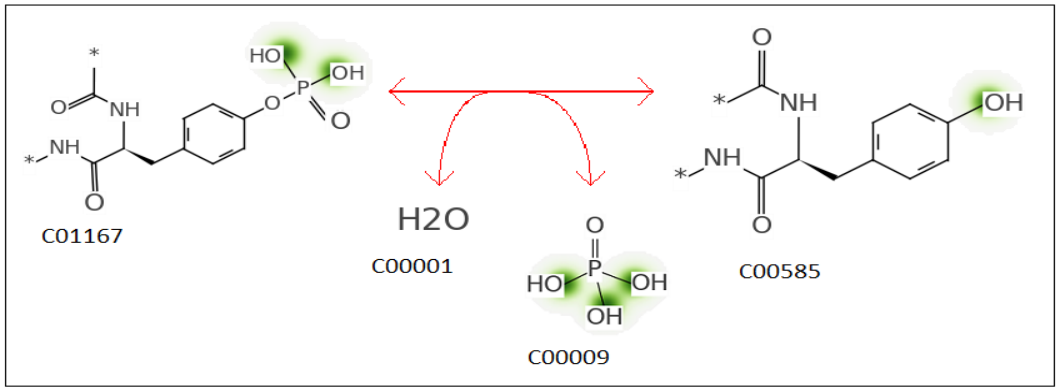
Slika 4.20: Prikaz kemijske reakcije R04535: C04619 <= >C05754 + C00001. Strukturna sprememba je v reakciji obarvana zeleno (SC(=O)CC).



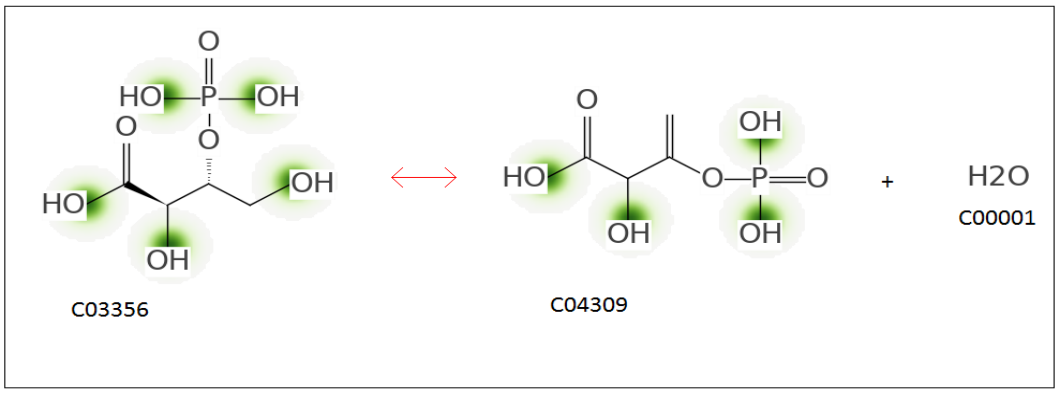
Slika 4.21: Prikaz kemijske reakcije R04965: C05757 <= >C05758 + C00001. Strukturna sprememba je v reakciji obarvana zeleno (SC(=O)CC).



Slika 4.22: Prikaz kemijske reakcije R01658: C00235 + C00129 <= >C00013 + C00341. Strukturna sprememba je v reakciji obarvana zeleno (OH).



Slika 4.23: Prikaz kemijske reakcije R02585: C01167 + C00001 <= >C00585 + C00009. Strukturna sprememba je v reakciji obarvana zeleno (OH).



Slika 4.24: Prikaz kemijske reakcije R04206: C03356 <= >C04309 + C00001. Strukturna sprememba je v reakciji obarvana zeleno (OH).

Poglavje 5

Zaključek

V diplomski nalogi smo z metodami strojnega učenja poskušali določiti pripadnost kemijskih reakcij ustreznim (binarnim) razredom. V prvem delu smo testirali različne nabore parametrov modeliranja na podatkih šestih osnovnih skupin kemijskih reakcij. Dosegli smo visoko vrednost AUC (0,97). V drugem smo poskušali s prej izbranim naborom najboljših parametrov določiti ustrezen binarni razred skupini kemijskih reakcij RNA-vezavnih proteinov. Končni dosežen rezultat je AUC 0,77. Na naključno izbrani skupini kemijskih reakcij s podobno sestavo osnovnih skupin kemijskih reakcij kot v skupini kemijskih reakcij RNA-vezavnih proteinov, je dosežen rezultat AUC 0,61, kar je za 0,16 slabši rezultat. Iz rezultatov lahko sklepamo, da imajo reakcije, v katere vstopajo RNA-vezavni proteini, svoje specifične lastnosti. Primerjava z rezultati testiranja osnovnih skupin kemijskih reakcij pa kaže na to, da so kemijske reakcije RNA-vezavnih proteinov bistveno bolj heterogene. Poiskali in prikazali smo pomembne vzorce, ki nastajajo kot produkti pri kemijskih reakcijah proteinov v interakciji z RNA. Napovedi modelov bi lahko izboljšali z dodajanjem in ovrednotenjem drugih načinov opisovanja kemijskih profilov (na primer MCS, MACCS keys). Vredno bi bilo preizkusiti ansambelske metode (zlaganje, angl. stacking) in globoke nevronske mreže. Smiselno bi bilo tudi razširiti nabor podatkov kemijskih reakcij.

Literatura

[1] R. C. Glen, A. Bender, C. H. Arnby, L. Carlson, S. Boyer, J. Smith, "Circular fingerprints: Flexible molecular descriptors with applications form physical chemistry to ADME" ,IDrugs. 2006 Mar;9(3):199-204.

[2] G. Landrum,"Large scale classification of chemical reactions from patent data", NIBR Informatics, Basel, Novartis Institutes for BioMedical, Research 10th International Conference on Chemical Structures, 10th German Conference on Chemoinformatics

[3] Mani, Inderjeet, and I. Zhang. "kNN approach to unbalanced data distributions: a case study involving information extraction."Proceedings of Workshop on Learning from Imbalanced Datasets. 2003.

[4] L. Bottou. "Large-scale machine learning with stochastic gradient descent."Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010. 177-186.

[5] M. Sharma, P. Garg, Čomputational Approaches for Enzyme Functional Class Prediction: A Review.Čurrent Proteomics 11.1 (2014): 17-22.

[6] P. D. Dobson, A. J. Doig, "Predicting enzyme class from protein structure without alignments."Journal of molecular biology 345.1 (2005): 187-199.

[7] T. Bray, A. J. Doig, J. Warwicker, Šequence and structural features of enzymes and their active sites by EC class."Journal of molecular biology 386.5 (2009): 1423-1436.

[8] L. De Ferrari, S. Aitken, J. van Hemert, I. Goryanin, "EnzML: multi-label prediction of enzyme classes using InterPro signatures."BMC bioinformatics 13.1 (2012): 61.

[9] V. Volpato, A. Adelifo, G. Pollastri, "Accurate prediction of protein enzymatic class by N-to-1 Neural Networks."BMC bioinformatics 14.Suppl 1 (2013): S11.."

[10] D. A. R. S Latino, J. Aires-de-Sousa, "Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests."Journal of chemical information and modeling 49.7 (2009): 1839-1846.

[11] M. Kotera, Y. Okuno, M. Hattori, S. Goto, M. Kanehisa, Čomputational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions."Journal of the American Chemical Society 126.50 (2004): 16487-16498.

[12] Y. Yamanishi, M. Hattori, M. Kotera, S. Goto, M. Kanehisa, "E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs."Bioinformatics 25.12 (2009): i179-i186.

[13] V. Egelhofer, I. Schomburg, D. Schomburg, "Automatic assignment of EC numbers."PLoS Comput Biol 6.1 (2010): e1000661.

[14] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckman, C. Strein, "Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.Čell 149.6 (2012): 1393-1406.

[15] J. Demšar, "Orange: data mining toolbox in Python."The Journal of Machine Learning Research 14.1 (2013): 2349-2353.

[16] G. E. A. P. A. Batista, R. C. Prati, M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data."ACM Sigkdd Explorations Newsletter 6.1 (2004): 20-29.

[17] J. Demsar, Štatistical comparisons of classifiers over multiple data sets."The Journal of Machine Learning Research 7 (2006): 1-30.

[18] M. Hentze, "An interview with Matthias Hentze."Trends in biochemical sciences 37.12 (2012): 507-508.

[19] W. Hentze, T. Preiss, "The REM phase of gene regulation."Trends in biochemical sciences 35.8 (2010): 423-426.

[20] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[21] KEGG. Dostopno na: http://www.kegg.jp/. Dostopano: maj 2015

[22] UnbalancedDataset. Dostopno na: https://github.com/fmfn/UnbalancedDataset. Dostopano junij 2015.

[23] RDKit, Open-Source Cheminformatics. Dostopno na: http://www.rdkit.org. Dostopano: maj 2015

[24] Morgan algorithm. Dostopno na: https://graphiteworks.wordpress.com/2011/08/31/chemoinformatics-curiosities-i-the-morgan-algorithm/. Dostopano: julij 2015

[25] Selecting good features part 3 - random forests. Dostopno na: http://blog.datadive.net/selecting-good-features-part-iii-random-forests/. Dostopano: julij 2015

[26] Naive Bayes. Dostopno na: http://scikit-learn.org/stable/modules/naive_bayes.html. Dostopano: julij 2015

[27] Support Vector Machine. Dostopno na: http://scikit-learn.org/stable/modules/svm.html. Dostopano: julij 2015

[28] Stohastic Gradient Descent. Dostopno na:
https://en.wikipedia.org/wiki/Stochastic_gradient_descent. Dostopano:
julij 2015

[29] Decision Tree Learning. Dostopno na:
https://en.wikipedia.org/wiki/Decisionlearning. Dostopano: julij 2015

[30] Molecular fingerprinting. Dostopno na:
http://www.daylight.com/dayhtml/doc/theory/theory.finger.html. Dostopano: junij 2015

[31] Creating molecular fingerprint. Dostopno na:
https://docs.chemaxon.com/display/CD/Chemical+Hashed+FingerpritnChemicalHashedFingerprint-fig1. Dostopano: avgust 2015

[32] Feature selection. Dostopno na:
https://en.wikipedia.org/wiki/Feature_selection. Dostopano: julij 2015

[33] Kemijska reakcija. Dostopno na:
https://sl.wikipedia.org/wiki/Kemijska_reakcija. Dostopano: julij 2015

[34] KEGG Wikipedia. Dostopno na:
https://en.wikipedia.org/wiki/KEGG. Dostopano: maj 2015

[35] Orange. Dostopno na:
http://orange.biolab.si/. Dostopano julij 2015.