Sveučilište Josipa Jurja Strossmayera u Osijeku
Sveučilište u Dubrovniku
Institut Ruđer Bošković
Poslijediplomski interdisciplinarni sveučilišni studij
Molekularne bioznanosti

Viktor Bojović

NOVI PARAMETRI ZA PROCJENU SLOŽENOSTI KLASIFIKACIJSKIH VARIJABLI I NJIHOVA PRIMJENA U MODELIRANJU SVOJSTAVA MOLEKULA

Doktorska disertacija

Osijek, 2020.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište Josipa Jurja Strossmayera u Osijeku **Doktorska disertacija**
Sveučilište u Dubrovniku
Institut Ruđer Bošković
Poslijediplomski interdisciplinarni sveučilišni
studij Molekularne bioznanosti, modul Bioinformatika

Znanstveno područje: Prirodne znanosti
Znanstvena polja: Kemija, Biologija

Novi parametri za procjenu složenosti klasifikacijskih varijabli i njihova primjena u modeliranju svojstava molekula.

Viktor Bojović

Disertacija je izrađena u: Centru za NMR, Institut Ruđer Bošković, Zagreb

Mentor: dr. sc. Bono Lučić, v. zn. sur. IRB u zn. zvanju zn. savjetnik, i nasl. docent Sveučilišta J. J. Strossmayera u Osijeku

Kratki sažetak doktorske disertacije:
Modeli koji opisuju odnos strukture i svojstva molekula trebaju sadržavati što manji broj strukturnih varijabli (molekulskih deskriptora) sa što većim sadržajem korisnih informacija. Informacijski sadržaj u strukturnoj varijabli povezan je s njenom složenošću i entropijom iskazanom brojem mogućih neidentičnih, nasumičnih realizacija varijable dobivenih permutiranjem njenih vrijednosti. Izvedeni su izrazi za izračun najmanje, najveće, prosječne nasumične vrijednosti i entropije parametara koji se koriste u procjeni kvalitete modela prilagođenih za analizu složenosti klasifikacijske varijable s dvije klase. Prikazana je i mogućnost poopćenja rezultata na druge vrste varijabli, te je razvijen mrežni poslužitelj za izračun složenosti skupa klasifikacijskih varijabli. Korisnost dobivenih rezultata prikazana je u primjenama na nekoliko skupova molekula u provjeri kvalitete varijabli i modela.

Broj stranica: 170
Broj slika: 37
Broj tablica: 27
Broj literaturnih navoda: 98
Jezik izvornika: hrvatski
Ključne riječi: QSAR modeliranje, klasifikacijske varijable, molekularni deskriptori, informacijski sadržaj, nasumična točnost, parametri kvalitete, složenost varijable, entropija varijable, server

Datum obrane: 14. prosinac 2020.

Stručno povjerenstvo za obranu:
1. izv. prof. dr. sc. Vesna Rastija, predsjednica povjerenstva
2. izv. prof. dr. sc. Domagoj Šimić, član
3. prof. dr. sc. Karolj Skala, član
4. izv. prof. dr. sc. Domagoj Matijević, zamjena

Disertacija je pohranjena u: Nacionalnoj i sveučilišnoj knjižnici Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb;Gradskoj i sveučilišnoj knjižnici Osijek, Europska avenija 24, Osijek; Sveučilištu Josipa Jurja Strossmayera u Osijeku, Trg sv. Trojstva 3, Osijek

BASIC DOCUMENTATION CARD

Josip Juraj Strossmayer University of Osijek **PhD thesis**
University of Dubrovnik
Ruđer Bošković Institute
University Postgraduate Interdisciplinary Doctoral Study of
Molecular biosciences, Bioinformatics module

Scientific Area: Natural sciences
Scientific Fields: Chemistry and Biology

Novel parameters for estimating the complexity of classification variables and their application in modeling properties of molecules

Viktor Bojović

Thesis performed at: NMR Centre, Ruđer Bošković Institute, Zagreb, Croatia

Supervisor: Dr. Bono Lučić, higher res. associate of RBI (in the sci. advisor rank)

Short abstract:
Structure-property relationship models should contain as few structural variables as possible (molecular descriptors) with as much useful information content as possible.The information content in a variable is related to its complexity and entropy represented by the number of possible nonidentical random realizations of variable obtained by permuting its values. Formulae are derived for minimal, maximal, mean random values and entropy of parameters used in the estimation of model quality and adapted for analysis of the complexity of two-state classification variables. The possibility of generalization of results to other types of variables is shown, and a web server is developed for estimation of the complexity of a set of classification variables. The usefulness of obtained results is demonstrated in analyzing the quality of variables and models.

Number of pages: 170
Number of figures: 37
Number of tables: 27
Number of references: 98
Original in: Croatian
Key words: QSAR modeling, classification variables, molecular descriptors, information content, random accuracy, quality parameters, variable complexity, entropy of variable, server

Date of the thesis defense: 14. December 2020.

Reviewers:
1. assoc. prof. dr. sc.Vesna Rastija, president of the committee
2. assoc. prof. dr.sc. Domagoj Šimić, committee member
3. prof. dr.sc. Karolj Skala, committee member
4. assoc. prof. dr. sc. Domagoj Matijević, substitute

Thesis deposited in: National and University Library in Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb; City and University Library of Osijek, Europska avenija 24, Osijek; Josip Juraj Strossmayer University of Osijek, Trg sv. Trojstva 3, Osijek

Doktorska disertacija izrađena je u Centru za nuklearnu magnetsku rezonanciju na Institutu Ruđer Bošković u Zagrebu, pod vodstvom dr. sc. Bone Lučića.

ZAHVALA

Ovaj rad ne bio moguć bez niza ljudi koji su to svojim djelovanjem na mene omogućili. Za rad najvažnija su bila znanja iz matematike, informatike, baza podataka i bioinformatike koja su se s vremenom slagala sloj po sloj do stupnja potrebnog za ostvarenje ovog cilja.

Za početke matematike zahvalan sam nastavnicama iz osnovne škole Lukreciji Protrki i pok. Ivanki Bronzović.

Za velik iskorak u znanju u matematici i vjeru u moja znanja vrlo sam zahvalan profesorici iz srednje škole pok. Vlatki Čulić.

Neizmjerno sa zahvalan Tomislavu Kalouseku, Marku Dučkiću i Bojanu Opačiću, Sanji Vladović i Vesni Grabić te doc. dr. sc. Panču Ristovu.

Vrlo sam zahvalan dr. sc. Sonji Nikolić s Instituta Ruđer Bošković (IRB) iz Zagreba i Igoru Petehu na prijateljskoj podršci, vjeri u mene i rješenju problema stanovanja u Zagrebu.

Neizmjerno sam zahvalan Sunčici Dodig na podršci i vjeri u mene sve ove godine.

Posebno sam zahvalan višem predavaču splitskog PMF-a Tončiju Dadiću za sva znanja iz baza podataka koja su me pratila kako izradom ove disertacije, tako i cijelom karijerom.

Veliko hvala profesorici sa splitskog PMF-a doc. dr. sc. Snježani Braić za sva znanja potrebna za izvođenje formula.

Zahvalan sam biofizičarima pokojnom doc. dr. sc. Damiru Zuciću sa Sveučilišta J. J. Strossmayera u Osijeku te emeritusu prof. dr. sc. Davoru Juretiću sa Sveučilišta u Splitu jer su mi znanja koja sam od njih dobio bila od velike važnosti za razumijevanje proteinskih struktura i razvijanje mojih sklonosti za istraživački rad.

Vrlo sam zahvalan dr. sc. Domagoju Kuiću na nesebičnoj pomoći pri razumijevanju fizikalnih pojmova koji su postali neizostavan dio ovog rada i brojnim diskusijama i pomoći kod učenja za ispit.

Velika hvala i kolegama s Instituta Ruđer Bošković iz Centru za NMR i Centru za informatiku i računarstvo i iz Zavoda za elektroniku, na konstruktivnim raspravama, savjetima, nesebičnoj podršci i pomoći, kao i djelatnicima stručnih službi koji su mi pomogli u svemu što je bilo potrebno.

Zahvalan sam bivšem i sadašnjem voditelju Centra za NMR - dr. sc. Draženu Vikiću-Topiću i dr. sc. Vilku Smrečkom, te voditeljici i zamjeniku voditeljice projekta „Bioprospekting Jadranskog mora“ na koji se vezalo moje istraživanje u postupku izrade doktorata - dr. sc. Rozelindri Čož-Rakovac i dr. sc. Matinu Roje na nesebičnoj pomoći, savjetima i kolegijalnoj, administrativnoj i organizacijskoj podršci.

Puno hvala mom mentoru dr. sc. Boni Lučiću, v. znan. sur. na brižnom mentorstvu, podršci, savjetima i velikoj pomoći oko provedbe istraživanja.

Vrlo sam zahvalan majci Jadranki, akademiku Nenadu Trinajstiću i Hrvatskoj Akademiji Znanosti i Umjetnosti za djelomično financiranje istraživanja u početnoj fazi moje istraživačke karijere.

Zahvaljujem se Ministarstvu Znanosti i Obrazovanja za financiranje dano Institutu Ruđer Bošković, te Hrvatskoj Vladi i Europskoj Uniji za financiranje kroz Europski fond za regionalni razvoji – Operativni program Konkurentnost i Kohezija (KK.01.1.1.01), Znanstveni centar izvrsnosti za bioprospekting mora– BioProCro. Ovaj doktorat u potpunosti je financiran od HRZZ od Europske Unije kroz Europski socijalni fond.

SADRŽAJ

1. UVOD

1.1. Početci modeliranja svojstava molekula

Istraživački rad u kemiji i bioznanostima pretežno je eksperimentalne prirode. Nakon što bi se prikupila dovoljna količina novih eksperimentalnih podataka, saznanja i informacija, pojedinci bi uspjeli uvidjeti određene jasne pravilnosti i napraviti sintezu postojećeg znanja vezanog za specifični problem. Takve sinteze znanja i generalizacije u svojoj biti imaju oblik modela, a bitno olakšavaju razumijevanje problema i ubrzavaju daljnji razvoj područja na koje se odnose. Prvom velikom sintezom kemijskoga znanja možemo smatrati pronalazak periodnoga sustava elemenata od strane ruskog kemičara Mendeljeyeva prije 150 godina, koja u osnovi ima svojstvo prvog generaliziranoga modela.

Prije nešto više od 50 godina radovima Corwina Hanscha u kemiji se počinju koristiti jednostavni multivarijatni modeli u svrhu pronalaženja struktura molekule koja bi mogle imati poboljšanu biološku aktivnost u odnosu na polazne poznate spojeve za koje su provedena eksperimentalne mjerenja aktivnosti [1,2]. Ti se modeli nazivaju modelima odnosa između strukture i svojstava ili aktivnosti molekula (*engl.* Quantitative Structure Property (Activity) Relationship, QSP(A)R, odnosno QSPR ili QSAR). Zbog uvriježenosti tih kratica izvedenih iz naziva ovih modela na engleskom jeziku u domaćoj znanstvenoj literaturi, bit će korištena kratica QSAR u tekstu disertacije.

QSAR modeli temelje se na pretpostavci da su svojstva (aktivnost) molekula određena njihovim strukturom, a ona je u vezi je s temeljnom pretpostavkom u biokemiji poznatom kao Anfinsenova hipoteza (pretpostavka) koja kaže da je biološka funkcija (i aktivnost) makromolekule (proteina ili nukleinskih kiselina) uglavnom određena njihovom primarnom strukturom [3]. Pritom se misli na male (organske) molekule koje imaju biološku aktivnost (npr. one molekule koje su lijekovi ili koje su potencijalni lijekovi).

Anfinsenova hipoteza potječe iz eksperimenta kojim je proučavano svijanje ribonukleaze A koristeći pri tome dvije molekule – ureu i beta-merkaptoetanol [3]. U tom se eksperimentu beta-merkaptoetanol koristio za smanjenje broja disulfidnih veza u proteinu dok se urea koristila za prekidanje nekovalentnih veza kao što su vodikove veze [4]. Posljedica upotrebe tih dviju molekula je denaturacija proteina zbog prekidanja prije svega kovalentnih disulfidnih, ali i vodikovih veza. Posljedica tog procesa je deaktivacija proteina, tj. prestanak njegove funkcije. Kako bi se funkcija vratila, potrebno je istovremeno odstraniti i ureu i beta-merkaptoetanol. Odstranjenjem samo jednog spoja, protein nije u mogućnosti vršiti svoju funkciju. Kad se odstrane oba spoja istovremeno protein se svije u odgovarajuće stanje slično početnom stanju. Iz toga je zaključeno da su za funkciju proteina nužne ispravna tercijarna (i sekundarna) struktura proteina, koja je određena njegovom (osnovnom) primarnom strukturom, tj. slijedom aminokiselina u proteinskoj sekvenci [3].

Iako prvobitno uvedena na primjeru proteina pokazujući izravnu vezu između strukture i funkcije proteina, Anfinsenova je hipoteza polazišna hipoteza i pri QSAR modeliranju malih molekula. Naime, također se pretpostavlja kako postoji uzročno-posljedična veza između strukture male molekule i njene (biološke) aktivnosti, ali i drugih njenih fizikalno-kemijskih svojstava. Odnos između strukture i aktivnosti nastoji se izraziti u obliku funkcionalne ovisnosti ili modela,

kako bi se lakše mogla pratiti promjena aktivnosti molekula kad se promijeni neki strukturni detalj u molekuli.

Polazišna točka u razvoju QSAR modela je kvantificiranje strukturnih svojstava molekula u obliku parametara (atributa, deskriptora) koji imaju kemijski jasno značenje koje se može definirati na skupu molekula na kojemu se provodi modeliranje. U prvim QSAR radovima deskriptori su fizikalno-kemijske prirode kao što su elektronski parametar koji ovisi o elektron-donorskim svojstvima karakterističnih kemijskih skupina u molekuli, koeficijent razdjeljenja (lipofilnost) te sterički parametar [1,2]. Međutim, u radu Free-ja i Wilsona deskriptori su indikatorske varijable (logički deskriptori) temeljeni na strukturi molekule, a definiraju se prema tome je li na nekom mjestu u molekuli supstituiran neki atom ili kemijska skupina (npr. CH_3, OH, NH, NH_2, ...) – vrijednost 1, ili nije supstituiran (vrijednost 0) [5]. Takve varijable nazivamo klasifikacijskim varijablama s dva stanja ili indikatorskim varijablama, tj. one indiciraju da je (ili nije) neka skupina supstituirana u strukturi na određenom mjestu.

Molekularni deskriptori računali su se u prošlosti (u vremenu početaka QSAR modeliranja) ručno, izvodeći i računajući svojstva molekule, ili uz pomoć priručnika ili knjiga s tablicama raznih svojstava. Tada se mogao izračunati samo mali skup molekularnih deskriptora. Računalnim programima danas je moguće računati na tisuće raznih molekularnih deskriptora, a za sve te izračune potrebno je imati strukturu molekule pohranjenu u elektroničkom obliku.

1.2. Elektronički zapis strukture i molekularni deskriptori

QSAR modeliranje danas se provodi računanjem velikoga broja molekularnih deskriptora pomoću računalnih programa razvijenih za tu svrhu (npr. program Dragon) [6]. S obzirom na to da je QSAR u pravilu multivarijatni model, kako bi dobiveni model bio pouzdan u statističkom smislu potrebno ga je razviti na (što) većem skupu molekula (primjera). Stoga, spomenute molekularne deskriptore potrebno je računati za veći skup srodnih molekula, što podrazumijeva u osnovi da se njihovo svojstvo ili aktivnost koje se želi modelirati:

(1) temelji na strukturama (tj. proizlazi iz struktura) molekula i

(2) ispoljava istim (ili jako srodnim, odnosno kompatibilnim) načinima djelovanja.

Međutim, kako bi bilo moguće računanje molekularnih deskriptora s tim programima, potrebno je molekulske strukture skupa molekula prikazati u elektroničkom obliku prikladnom za računalnu obradu. Elektronički oblik strukture, spremljen u obliku datoteke, mora biti organiziran prema prihvaćenim kemijskim standardima kako bi ga mogli jednoznačno koristiti korisnici raznih računalnih programa u području modeliranja. Razvijeni standardni oblici datoteka, koji sadrže informacije o strukturi molekula, moraju biti prikladni za učitavanje i spremanje u računalnu memoriju.

1.2.1. SMILES oblik strukture

Za potrebe vizualizacije na računalu, simulacija ili izračuna molekularnih deskriptora), razvijen je poseban skup pravila kojima se zapisuju sve važne strukturne informacije u elektroničkom obliku (formatu) u jednoj datoteci. Najjednostavniji takav zapis strukture molekula naziva se SMILES (*engl.* Simplified Molecular-Input Line-Entry System), definiran kao jezik za specijalnu upotrebu u kemiji kako bi opisao prirodu i topologiju molekularne strukture [7,8].

S pomoću skupa strukturnih pravila SMILES, struktura molekule zapisuje se slijedeći (kao vodilje) kovalentne kemijske veze u definiranom redoslijedu**,** pri čemu se atom po atom zapisuju onim redom kojim su povezani u strukturi. Dakle, za svaki atom zapisuju se susjedni atomi, a za susjedne atome njihove susjedne atome i tako redom. Pritom se vodik uglavnom ne zapisuje nego se njegova prisutnost u organskim spojevima podrazumijeva. Iz tog se razloga metan zapisuje kao 'C', etan 'CC', a propan 'CCC', pri čemu crtica (ukoliko je uključena u zapisu) označava da su susjedni atomi povezani jednostrukom kovalentnom vezom. U SMILES strukturi amonijaka zapisuje se samo dušik 'N', umjesto klasičnog zapisa koji bi bio 'NH_3'. U slučaju iona, koriste se uglate zagrade pa je zlato [Au], a proton je [H+]. Dvostruka kemijska veza zapisuje se pomoću dviju crtica, tj. znaka jednakosti '=', a trostruka pomoću znaka '#'. Za slučaj cikličkih struktura, početni i krajnji atom u prstenu označi se istim brojem označavajući tako početak i završetak prstena. Tako se SMILES struktura ciklopropana zapisuje kao 'C1CC1'.

U slučaju razgranavanja koriste se i zagrade pa se 3-etil-amin zapisuje u obliku CCN(CC)CC [9]. Strukturu svake molekule može se početi pisati od bilo kojeg atoma, te se tako ista struktura može zapisati na više načina. U slučaju potrebe za vizualizacijom SMILES formata, moguće je koristiti komercijalne ili besplatne servise dostupne putem Interneta (kao što je npr. poslužitelj Online Smiles Translator) [10-12]. SMILES format zapisuje samo atome i veze među njima, ali najčešće ne i koordinate atoma.

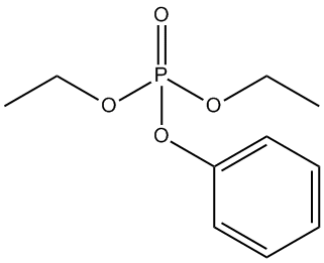
1.2.2. Zapis strukture u obliku MOL/SDF datoteke

Osim SMILES oblika (formata) strukture, poznati su još CIF, PDB, MOL ili SDF (*engl.* SDF - Structure Data File) i brojni drugi. U ovoj disertaciji bit će korišten i ukratko opisan samo MOL/SDF oblik, gdje osnovni oblik zapisa strukture jedne molekule ima nastavak (ekstenziju) MOL (*engl.* MOLecule), a više takvih struktura spremljenih u jednoj datoteci ima nastavak SDF. Za zapis koordinata atoma koriste se CIF i PDB formati koji su prilagođeni za pohranu eksperimentalnih podataka jer pohranjuju i dodatne informacije o eksperimentalnoj metodi, alternativnim lokacijama atoma, pripadnost lancu, ligandima itd. CIF i PDB strukturirane datoteke prilagođene su za makromolekule, dok je za manje molekule uglavnom u uporabi MOL/SDF oblik datoteka.

Datoteke u MOL obliku pohranjuju informacije o kiralnosti, koordinatama atoma te njihovim vezama, uključujući njihove duljine, a zapis strukture završava s 'M END'. Datoteka SDF u osnovnom obliku sadrži više struktura pohranjenih u MOL formatu, jednu ispod druge, odijeljene znakom '$$$$'. SDF datoteka može sadržavati i dodatne informacije poput naziva spoja, jedne ili više inačica SMILES strukture molekule, CAS broja, te fizikalno-kemijskih svojstava molekula spremljenih u standardiziranome obliku (u dva retka) [13].

Kako svaki datotečni format (zapis strukture) nije čitljiv u svim programima, ponekad je potrebno raditi pretvorbe jednog oblika zapisa (datoteke) u drugi (ili jedne datoteke u drugu). Tu pretvorbu moguće je napraviti pomoću besplatnih programa poput programa Openbabel koji se može koristiti on-line ili preko stranice Cheminfo [14,15].

Za ilustraciju razlika među zapisima datoteka prikazan je primjer strukture kemijskoga spoja dietil-fenil-fosfata iz rada Hanscha i Fujite (u kojem je osnovni spoj redni broj 8 u *Tablici 3*). SMILES oblik toga spoja je CCO[P](=O)(OCC)OC1=CC=CC=C1, a struktura je prikazana na *Slici 1.1* [1]. Datoteka te strukture u obliku MOL/SDF dana je u *Prilogu* 1.1. (Zapis strukture dietil-fenil-fosfata sa *Slike 1.1* u obliku MOL/SDF).

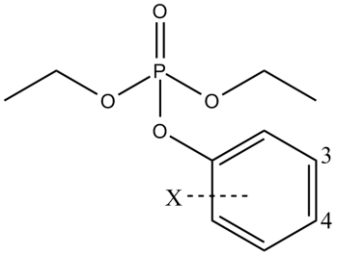


Slika 1.1 Osnovni oblik strukture dietil-fenil-fosfata iz ref. [1]

Unutar SMILES strukture aromatski prsten zapisan je kao 'C1=CC=CC=C1', te je vidljivo da je vezan na kisik (O) koji je vezan na fosfor [P]. Taj fosfor povezan je s ostatkom strukture s ukupno četiri veze.

1) CCO - jednostruka veza
2) (=O) – dvostruka veza
3) (OCC) - jednostruka veza
4) jednostruka veza s O iz preostalog dijela strukture (OC1=CC=CC=C1)

Na *Slici 1.2* prikazan je općeniti oblik strukture dietil-fenil-fosfata iz *Tablice 3* iz rada Hanscha i Fujite [1]. U toj tablici u [1] osnovna struktura nalazi se pod rednim brojem 8.



Slika 1.2. Općenita struktura dietil-fenil-fosfata iz koje su sintetizirani spojevi u *Tablici 3* iz ref. [1]

Zajednički dio strukture za sve spojeve sa *Slike 1.2* koji ima vodikov atom na četiri slobodna mjesta u aromatskom prstenu. U radu Hanscha i Fujite taj osnovni oblik strukture modificiran je u više navrata dodavanjem raznih kemijskih skupina (X) na jedno od dva mjesta označena u aromatskom prstenu brojevima 3 ili 4 [1]. Tako je dobiven skup od ukupno 14 spojeva iz klase dietil-fenil-fosfata na kojem su razvijani regresijski QSAR modeli toksičnosti tih spojeva na kućne muhe. Stupanj toksičnosti iskazan je mjerenjima koncentracija (C) svakog od 14 ispitivanih spojeva pri kojoj je 50 % organizama uginulo i, iz praktičnih razloga, iskazan je kao $log(1/C)$. Modeli su temeljeni samo na dva fizikalno-kemijska parametra: σ - elektronski parametar koji opisuje specifičnosti raspodjele naboja na molekuli i π - particijski koeficijent za sustav oktanol-voda koji opisuje lipofilno-hidrofilna svojstva molekule, a ona su važna za prolazak kroz staničnu membranu, što je osnovni preduvjet za ispoljavanje biološke aktivnosti nekog kemijskog spoja [1]. Tako prvi oblik QSAR modela ima oblik multivarijatne regresijske jednadžbe:

$$log(1/C) = a\pi + b\sigma + c \qquad (1.1)$$

gdje su a, b i c (konstanta) optimalni parametri koji se dobivaju optimizacijskim algoritmom (metoda najmanjih kvadrata). U tom postupku određivanja optimalnih regresijskih parametara minimizira se srednje kvadratno odstupanje između eksperimentalnih vrijednosti $log(1/C)$ i vrijednosti dobivenih iz jednadžbe (QSAR modela) (1.1).

1.2.3. Molekularni strukturni deskriptori i njihov izračun

Iz elektroničkoga zapisa strukture računaju se dostupnim računalnim programima numeričke varijable (deskriptori, X-varijable) koje kvantificiraju strukturne karakteristike i posebnosti svake molekule. U kemiji je molekularni deskriptor završni rezultat logičke i matematičke procedure koji transformira kemijsku informaciju iz strukture molekule u brojčanu vrijednost [16]. Taj broj daje uvid u fizikalno-kemijska, konstitucijska, topološka ili druga izvedena svojstava molekula. Primjer jako često korištenog računalnog programa je Dragon [6], koji je razvijen u grupi prof. Todeschinija iz Milana. Taj je program unatrag 20 godina nadograđivan kroz više inačica, a trenutačna 7.0 računa preko 5000 molekularnih deskriptora iz različitih klasa. Najranije su u QSAR modeliranju korištene klase konstitucijskih deskriptora. Primjeri takvih deskriptora su molekulska težina, broj atoma, broj atoma određene vrste (npr. broj ugljikovih atoma), broj određenih funkcionalnih skupina (npr. OH ili CH_3 skupina), broj dvostrukih veza, itd. Razlog je jasan - od početaka razvoja QSAR modeliranja prije 55 godina pa do unatrag 25 godina računalna tehnologija bila je vrlo slabo razvijena, elektronički zapisi strukture nisu bili definirani i standardizirani, i nije bilo moguće računati složenije molekulske deskriptore za veće skupove molekula.

Mali broj fizikalno-kemijskih deskriptora, poput ranije spomenutih particijskih koeficijenata π ili elektronski parametar σ bili su izračunani za različite funkcionalne skupine, i objavljene u radovima, knjigama ili u specijaliziranim priručnicima. Izračun takvoga parametra se za neku molekulu provodio pregledom odgovarajuće literature i zbrajanjem doprinosa svih skupina u molekuli. Takav postupak bio je vremenski dugotrajan, često povezan s računskim/računalnim pogreškama i rijetko je bilo moguće raditi QSAR modele za veće skupove molekula (npr. > 50 ili 100 molekula). Zagrebačka grupa, vođena I. Gutmanom i N. Trinajstićem uvela je 1972. godine topološke deskriptore, $M1$ i $M2$, kasnije nazvane zagrebački indeksi [17]. Vrlo je često prisutna u QSAR modelima i nadogradnja ta dva topološka deskriptora nazvana indeks povezanosti [18].

Topološki deskriptori računaju se iz grafa koji reprezentira kemijsku strukturu molekule na način da vrhovi predstavljaju atome, a veze među atomima predstavlja grane grafa. Topološki deskriptori računaju se matematičkim postupcima (algoritmima) iz matričnoga prikaza grafa i na temelju svojstava grafa (koji je pojednostavljena struktura molekule). Najčešće korišteni matrični prikazi svojstava grafa su matrica susjedstva čvorova, matrica povezanosti među čvorovima grafa ili matrica šetnji (duljine jedne ili više veza) na grafu [19].

Pored spomenutih skupina deskriptora razvijene su i brojne druge poput fizikalno-kemijske skupine. Ti se deskriptori temelje na elektronskoj strukturi, popunjenosti orbitala koje odgovaraju reaktivnim elektronima, energiji molekule, raspodjeli naboja, nabijenoj površini pristupačnoj otapalu, itd. Fizikalno-kemijski deskriptori računaju se semi-empirijskim ili kvantno-kemijskim postupcima s pomoću raspoloživih programa poput programa MOPAC ili metode DFT (*engl.* Density Functional Theory) implementirane u programskom paketu Gaussian [20,21].

1.2.4. Indikatorski deskriptori i deskriptori „otiska prsta“ (*engl. fingerprints*) kao posebna skupina klasifikacijskih varijabli

Danas vrlo često korišten pristup, koji se u medicinsko-kemijskoj literaturi počesto naziva i Free-Wilsonov pristup, temelji se na deskriptorima koji poput „otiska prsta“ (engl. fingerprint“) opisuju molekulu Ta vrsta deskriptora reprezentira strukturu molekule prema: (1) svim mogućim supstitucijskim mjestima u osnovnoj strukturi (kosturu) koji je zajednički cijelom skupu molekula, i (2) po svim prihvatljivim supstituentima (za tu vrstu molekula). Oba uvjeta iz prethodne rečenice

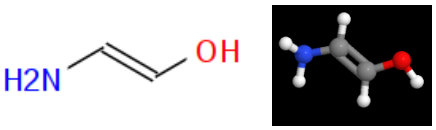
odnose se na potrebu da molekule određene klase, koje dijele strukturnu sličnost (tj. imaju zajednički dio strukture - kostur), ispoljavaju svoju aktivnost ili svojstvo istim mehanizmima, što je nužni preduvjet kako bi se na neki skup molekula mogli primijeniti QSAR modeli [16].

U analizi podataka, najelegantniji način njihovog prikazivanja je tablica, gdje stupci tablice sadrže vrijednosti varijabli (deskriptora), a redci informacije o vrijednostima raznih varijabli za isti slučaj (tj. molekulu) u skupu podataka. Same vrijednosti varijable mogu biti cijeli, ili realni (decimalni) broj, ili u nekim slučajevima slovčana oznaka klase, itd. U ovom radu cilj je procijeniti (i numerički iskazati) složenost varijabli u slučaju kada se njene numeričke vrijednosti sastoje od samo dvije vrijednosti (0 i 1), koje označavaju pripadnost molekule jednoj od dvije klase. Varijabla koja sadrži informaciju o tome ima li nekog svojstva (npr. kemijske skupine) u molekuli ili nema, naziva se indikatorskom varijablom.

U početnoj fazi modeliranja često se provodi manje strogi postupak isključivanja iz daljnjih analiza onih varijabli koje ne ispunjavaju osnovne preduvjete varijabilnosti i informativnosti [23]. U QSAR modeliranju složenijih aktivnosti i na velikim skupovima molekula nije moguće dobiti zadovoljavajući model sa samo jednim deskriptorom. Stoga se za velike skupove u pravilu razvijaju modeli multivarijatni modeli ili modeli s većim brojem deskriptora. Deskriptori sa slabom prediktivnom moći ponekad se zadržavaju u modelima zbog njihove teorijske pozadine i mogućnosti interpretacije. Međutim, nije istraženo, niti precizirano u literaturi koji bi bili minimalni uvjeti kako bi neki deskriptor sadržavao minimalnu količinu korisne informacije te to će biti tema istraživanja u disertaciji. Broj deskriptora koji se danas može izračunati za neki skup molekula u pravilu je jako velik [24]. Tako postupci izbora malog podskupa značajnih (signifikantnih) deskriptora postaju sve teži i računalno (algoritamski) sve zahtjevniji. S druge pak strane, povećanjem broja deskriptora u modelu kod velikih skupova podataka (molekula) povećava se mogućnost preciznijeg objašnjenja aktivnosti molekule, jer jedan deskriptor ne može objasniti svu složenost i varijabilnost aktivnosti molekula. Problem eliminacije nesignifikantnih varijabli (deskriptora) rješava se najprije pre-selekcijom, gdje se deskriptori s niskom varijancom odbacuju iz analiza. Potom se koriste različiti algoritmi za selekciju varijabli, a među prvim uspješnim primjenama ističu se genetski algoritmi [25]. Međutim, danas se za manje skupove deskriptora može provesti i pretraživanje svih mogućih kombinacija (podskupova) deskriptora u modelima, te se na koncu može izabrati najbolji model, prema jednom ili više odabranih parametara kvalitete [26].

Pored početnih linearnih i jednostavnih nelinearnih multivarijatnih QSAR modela [1], kasnije su razvijeni postupci za nelinearne QSAR modele poput neuronskih mreža [27]. Ti se modeli, kao i svaki općeniti nelinearni model, mogu aproksimirati razvojem funkcije u red potencija po nezavisnim ulaznim varijablama (molekularnim deskriptorima) [26]. Primjena tih algoritama nije moguća bez prethodne selekcije varijabli, koja se provodi izvan algoritma neuronskih mreža (npr. genetskim algoritmima [25]), a najčešće se kao ulaz u modele neuronskih mreža koriste deskriptori izabrani kao najbolji u multivarijatne regresijske modele [27].

Za lakšu identifikaciju molekula, njihovo ubrzano pretraživanje i lakši pronalazak sličnih molekula u bazama služe strukturni fingerprintovi, tj. nizovi informacija o strukturi gdje svaki bit u tom velikom nizu bitova ne mora imati neko jasno definirano strukturno značenje, a koristan je kod analize sličnosti među molekulama [28]. Strukturni fingerprint deskriptori mogu sadržavati informacije o atomu, susjedima atoma, grupama veza ili npr. stazama veza različitih duljina. Kao primjer kreiranja strukturnog fingerprinta bit će korištena molekula 2-aminoetenola čija je SMILES formula OC=CN, a njen izgled prikazan je na *Slici 1.3*.



Slika 1.3 prikaz strukture 2-aminoetenola

Kako bi se dobio strukturni fingerprint, algoritmi koji ga računaju rade ekstrakciju staza koje se sastoje od atoma (čvorova) i veza. Za stazu duljine 0 uključuju se samo atomi, za duljinu 1 dva atoma i veza između njih, za duljinu 3 uključena su 3 atoma i veze između njih, itd. Primjer takvih staza za 2-aminoetenol je slijedeća: C, O, N, OC, C=C, CN, OC=C, C=CN, OC=CN. Takve staze potrebne su za stvaranje strukturnih fingerprinta.

Ukoliko uzorak koji tražimo postoji kao dio neke molekule u nekoj bazi, tada će se taj dio strukturnog fingerprinta poklopiti, što olakšava pronalaženje. Osim informacija o molekuli, strukturni fingerprint može sadržavati i informacije o reakciji, pa se tako mogu spremati informacije o razlici između reaktanta i produkta. Ukoliko je potrebno usporediti dvije molekule po sličnosti, to je moguće učiniti pomoću strukturnog fingerprinta koristeći razne algoritme za određivanje sličnosti, primjerice Tanimotov algoritam [29].

U istraživanju lijekova uvodi se pojam farmakofora koja je definirana kao skup kemijskih svojstava (funkcionalnih grupa, npr. –OH ili $–CH_3$ skupina) i njihovo prostorno uređenje koje definira farmakološku specifičnost za skupine kemijskih spojeva [30]. Primjer takvih podataka,organiziranih u obliku varijabli koje sadrže informaciju o strukturnim obilježjima skupa spojeva i koje se nazivaju indikatorskim varijablama, prikazan je u dvije tablice u nastavku.

Definiranje indikatorskih varijabli u kemijskom modeliranju ilustrirat će se na primjeru skupa spojeva preuzetog iz rada Hanscha i Fujite [1], koji se ubraja u prve radove iz područja QSAR modeliranja, a sam Corwin Hansch smatra se utemeljiteljem tog područja. Time se želi pokazati da su indikatorski deskriptori bili uvedeni među prvima u modeliranje u kemiji. Nadalje, ti se deskriptori koriste neprekidno do današnjih dana, a predstavljaju klasifikacijske varijable s dva stanja na koje će se moći izravno primijeniti parametri za izračun složenosti varijabli koji su predmet istraživanja u ovoj disertaciji. Uporaba indikatorskih deskriptora u QSAR modeliranju od samih početaka ima svoje jasne i opravdane razloge. Naime, početni QSAR modeli razvijeni su u području sintetske medicinske kemije kada su eksperimentalni sintetski kemičari provodili modifikacije kemijskih spojeva (lijekova) s ciljem dobivanja novog kemijskog spoja poboljšanih svojstava. U tim istraživanjima, posao bi im olakšao QSAR model s jednostavnim deskriptorima (kakvi su indikatorski deskriptori) koji pomažu u optimizaciji postupaka modifikacije spojeva. Takvim postupkom optimizacije štedi se ljudski rad, ali se štedi i na potrošnji kemikalija i drugih resursa potrebnih za provedbu sinteze kemijskih spojeva. Naime, indikatorska varijabla vezana je uz jedno zajedničko mjesto u svim spojevima iz skupa molekula, i na njemu se svaki put ugradi u strukturu nova kemijska skupina koja ima svoja specifična svojstva. Na taj način, važnost nekog indikatorskog deskriptora izravno upućuje na to koja vrsta kemijskih skupina ugrađena u strukturi na točno definirano mjesto daje spoj s najboljim željenim svojstvom ili aktivnošću. Na primjer, ako se želi pronaći spoj s poboljšanim svojstvom inhibicije nekog enzima (proteina) važnog u razvoju neke bolesti, onda će razvijeni QSAR model dati informaciju o tome koje kemijske skupine ugrađene u osnovnu strukturu daju spoj koji je najučinkovitiji inhibitor promatranog enzima.

U radu [1] pretpostavljeno je da se promjena elektronskog svojstva derivata osnovnog (roditeljskog) spoja (*Slika 1.1*) događa isključivo zbog razlike u vrijednosti σ funkcionalne skupine na aromatskom prstenu na položajima 3 ili 4 (*Slika 1.2*) koja karakterizira taj spoj. Analogno vrijedi za promjenu lipofilnosti spojeva iskazanu vrijednostima deskriptora π, koja se također numerički iskazuje u odnosu na roditeljski spoj (*Slika 1.1*) za koji su vrijednosti σ i π jednake nuli (*Tablica 1.1*). Svi supstituenti kovalentno se vežu za dietil-fenil-fosfat, i to za dio označen s X na *Slici 1.2.*

Tablica 1.1 Primjer definiranja indikatorskih varijabli. Skup molekula za QSAR modeliranje sastavljen je prema *Tablici 3* iz rada [1]

#	X [a]	P.P.[b]	σ	π	I_3	I_4	SMILES cijeloga spoja
1	NO_2	4	1.27	0.46	0	1	CCO[P](=O)(OCC)OC1=CC(=CC=C1)N(O)O
2	SO_2CH_3	4	1.05	-0.03	0	1	CCO[P](=O)(OCC)OC1=CC=C(C=C1)S(=O)(=O)C
3	CN	4	1	0.1	0	1	CCO[P](=O)(OCC)OC1=CC=C(C=C1)CN
4	NO_2	3	0.71	0.71	1	0	CCO[P](=O)(OCC)OC1=CC=C(C=C1)N(O)O
5	SF_6	3	0.68	1.92	1	0	CCO[P](=O)(OCC)OC1=CC(=CC=C1)[S](F)(F)(F)(F)(F)F
6	Cl	4	0.23	0.89	0	1	CCO[P](=O)(OCC)OC1=CC=C(C=C1)Cl
7	t-Bu	3	-0.12	1.6	1	0	CCO[P](=O)(OCC)OC1=CC(=CC=C1)C(C)(C)C
8	H	-	0	0	0	0	CCO[P](=O)(OCC)OC1=CC=CC=C1
9	$N(CH_3)_2$	3	-0.21	0.06	1	0	CCO[P](=O)(OCC)OC1=CC(=CC=C1)N(C)C
10	COOH	4	0.73	0.08	0	1	CCO[P](=O)(OCC)OC1=CC=C(C=C1)C(=O)O
11	t-Bu	4	-0.2	1.55	0	1	CCO[P](=O)(OCC)OC1=CC=C(C=C1)C(C)(C)C
12	OCH_3	3	0.12	0.08	1	0	CCO[P](=O)(OCC)OC1=CC(=CC=C1)OC
13	OCH_3	4	-0.27	-0.16	0	1	CCO[P](=O)(OCC)OC1=CC(=CC=C1)OC
14	CH3	4	-0.17	0.44	0	1	CCO[P](=O)(OCC)OC1=CC=C(C=C1)C

[a] Kemijska vrsta supstituenta (X) prema slici 1.2. [b] Odnosi se na položaj supstituenta X u prstenu, prema slici 1.2.

Međutim, pritom u spojevima 1-7 i 9-14 u *Tablici 1.1* nije uzeta u obzir moguća interakcija supstituenata međusobno i s ostatkom molekule, jer manje kemijske skupine na mjestima 3 i 4 na *Slici 1.*2 u strukturi roditeljskog spoja (*Slika 1.1*) imat će slabiju interakciju s ostatkom molekule nego veći supstituenti (funkcionalne skupine na mjestima 3 i 4)). Stoga, uvedene su dvije indikatorske varijable (deskriptori) kako bi se utvrdio utjecaj međusobnih interakcija supstituenata na položajima 3 i 4 (*Slika 1.2*) na biološku aktivnost spojeva. Te dvije indikatorske varijable (deskriptori) označeni su u *Tablici 1.1* s I_3 i I_4, i imaju vrijednost 1 kad je supstituent prisutan na položaju 3 odnosno 4, a inače im je vrijednost jednaka 0.

Ukoliko se podaci oblikuju na način da se na položajima 3 i 4 promatra postojanje točno određene funkcionalne skupine (supstituenta), jasno je da broj indikatorskih varijabli raste, što se vidi iz *Tablice 1.2.* U *Tablici 1.2*, informacije su razdijeljene u više varijabli koje imaju manje informacija nego u *Tablici 1.1*, a vezane su samo za utvrđivanje prisustva (1) ili odsustva (0) jedne kemijske skupine na točno određenom mjestu u skupu molekula.

Tablica 1.2 Primjer definiranja indikatorskih varijabli za strojno učenje. Skup molekula za QSAR modeliranje sastavljen je prema *Tablici 3* iz rada [1] [a]

#	σ	π	P.P. [b]	Supstituent											
				NO_2	SO_2CH_3	CN	NO_2	SF_6	Cl	t-Bu	H	$N(CH_3)_2$	COOH	OCH_3	CH_3
1	1.27	0.46	4	1	0	0	0	0	0	0	0	0	0	0	0
2	1.05	-0.03	4	0	1	0	0	0	0	0	0	0	0	0	0
3	1	0.1	4	0	0	1	0	0	0	0	0	0	0	0	0
4	0.71	0.71	3	1	0	0	0	0	0	0	0	0	0	0	0
5	0.68	1.92	3	0	0	0	0	1	0	0	0	0	0	0	0
6	0.23	0.89	4	0	0	0	0	0	1	0	0	0	0	0	0
7	-0.12	1.6	3	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	-1	0	0	0	0	0	0	0	1	0	0	0	0
9	-0.21	0.06	3	0	0	0	0	0	0	0	0	1	0	0	0
10	0.73	0.08	4	0	0	0	0	0	0	0	0	0	1	0	0
11	-0.2	1.55	4	0	0	0	0	0	0	1	0	0	0	0	0
12	0.12	0.08	3	0	0	0	0	0	0	0	0	0	0	1	0
13	-0.27	-0.16	4	0	0	0	0	0	0	0	0	0	0	1	0
14	-0.17	0.44	4	0	0	0	0	0	0	0	0	0	0	0	1

[a] Oznake kratica objašnjenje se u tekstu i *Tablici 1.1*. [b] P.P. odnosi se na položaj supstituenta X u prstenu, prema *Slici 1.2*.

Međutim, iako su takve varijable daleko monotonije (vrijednosti su im degenerirane, tj. imaju više identičnih vrijednosti) nego izvorne indikatorske varijable iz *Tablice 1.1*, one su prikladnije za računalnu obradu i za metode strojnog učenja koje se danas sve češće koriste u QSAR modeliranju. Takve varijable su u biti podskup današnjih skupova strukturnih fingerprint deskriptora, koji se redovito koriste u QSAR modeliranju u farmaceutskim istraživanjima razvoja novih lijekova. Uporaba velikog broja takvih deskriptora u modeliranju čini modele previše složenim, a u statističkom smislu i nedovoljno pouzdanim jer se, za određeni broj molekula koji je stalan i unaprijed definiran, u model uključuje bitno veći broj deskriptora. Drugi je problem što svaki od tih uključenih deskriptora (varijabli) ima jako degenerirane vrijednostima, tj. prevladavaju vrijednosti jedne klase (ili vrijednosti 0, ili vrijednosti 1). Upravo se istraživanje u disertaciji s ciljem definiranja i izvođenja parametara za izračun i kvantificiranje složenosti pojedinačnih klasifikacijskih varijabli odnosi na takve varijable (deskriptore). Izvedeni parametri poslužit će u ranoj detekciju deskriptora koji su niske složenosti, što će omogućiti racionalnu i učinkovitu prethodnu selekciju i isključivanje takvih deskriptora (varijabli) iz daljnjeg postupka modeliranja.

1.3. Procjena kvalitete QSAR modela i nasumične korelacije/točnosti

U prvim godinama nakon uvođenja metode QSAR računao se mali broj molekularnih deskriptora, ali su bili jako mali i skupovi molekula na kojima su se modeli razvijali. Stoga, već u to vrijeme uočio se problem nasumične korelacije među podacima [31,32], te se simulacijama došlo do zaključka kako broj varijabli u multivarijatnim modelima ne smije biti veći od 1/5 broja molekula u skupu na kojem se model razvija [32]. Na razini EU, QSAR modeli koriste se u procjeni štetnog utjecaja kemijskih spojeva na okoliš i žive organizme, te su razvijeni postupci za procjenu njihove kvalitete [16]. Točnost klasifikacijskog modela s dva stanja (dvije klase, 0 i 1) čije su predviđene

vrijednosti označene kao varijabla X u odnosu na eksperimentalne vrijednosti klasifikacijske varijable s dva stanja Y na kojima je razvijen može se iskazati (izračunati) raznim parametrima za izračun točnosti klasifikacijskog modela [33]. Samo neki od tih parametara preporučuju se i u dokumentu OECD ((*engl*. Organisation for Economic Co-operation and Development, Organizacija za ekonomsku suradnju i razvoj) vezano uz postupke koji se provode u validaciji klasifikacijskih QSAR modela razvijenih za primjenu u zaštiti okoliša i zdravlja [16]. Taj dokument postao je sastavni dio normi EU i Hrvatske, a sastavila ga je i 2007. godine objavila skupina eksperata (znanstvenika) nakon višegodišnjega rada na standardizaciji i validaciji QSAR modela razvijenih na podacima važnim u zaštiti okoliša i zdravlja. Kriteriji koje mora ispunjavati model kako bi bio primjenjiv u zaštiti okoliša, definirani su nakon opsežne analize rezultata najznačajnijih istraživanja iz područja QSAR modeliranja koji su do tada bili poznati.

Iskazivanje stvarne točnosti klasifikacijskih modela, kao i njihovo rangiranje prema mjerama (parametrima) kvalitete, iznimno je važno u brojnim istraživačkim područjima. Za potrebe istraživanja moguće je za svake binarne podatke odrediti njihov nasumični model permutacijama i izračunati njegovu najvjerojatniju nasumičnu točnost te ju usporediti sa stvarnom točnošću [34]. Pojam *„najvjerojatnija točnost"* u ovom radu se mijenja u pojam „*Prosječna nasumična točnost*". Razlog tomu je što postoje slučajevi kada vrijednost nasumične točnosti ne postoji među vrijednostima stvarne točnosti, a uočeno je pravilo da se taj parametar uvijek odnosi na srednju vrijednost prilikom testiranja na velikom uzorku. Isto pravilo vrijedi i za druge parametre u ovom radu.

Razlika između stvarne točnosti odabranog modela i njegove prosječne nasumične točnosti pokazatelj je težine pojave/problema koji se modelira kao i kvalitete samoga modela (označen kao ΔQ_2 u [34,35]). Taj rezultat prikazan na modeliranju sekundarne strukture proteina [34] bio je poticaj za proširenjem istraživanja i na indikatorske varijable (i strukturne fingerprint deskriptore) u QSAR modelima.

Uravnoteženi model definiran u [34,35] je onaj model koji predviđa brojeve elemenata (slučajeva) u klasama jednake onim u eksperimentalnoj varijabli. To vrijedi bez obzira radi li se o varijabli s dvije ili više klasa. Međutim, planirana istraživanja u disertaciji odnose se na klasifikacijske varijable s dva stanja. To je najjednostavniji slučaj klasifikacijske varijable, a ujedno je u zadnje vrijeme i sve češće korišten u QSAR modeliranju. Taj trend potaknut je sve većom digitalizacijom znanja, informacija i podataka, pa tako i digitalizacijom informacije sadržane u strukturama molekula koja se kodira u obliku deskriptora kakvi su indikatorski deskriptori i strukturni fingerprintsi opisani ranije.

1.4. Svrha i cilj istraživanja

Ideja za istraživanje izloženo u ovoj disertaciji temelji se na konceptu nasumičnog modela koji je uveden ranije na primjeru klasifikacijskog modela s dva stanja [34]. Pritom je izveden izraz za nasumičnu točnost ($Q_{2,rnd}$) [34,35]

$$Q_{2,rnd} = \frac{p+u \quad p+o \quad +(n+o)(n+u)}{N^2}. \tag{1.2}$$

koja je originalno nazvana „najvjerojatnijom nasumičnom točnošću". Jednadžba (1.2) opisuje nasumičnu točnost klasifikacijskog modela s dva stanja koji predviđa ($p + o$) vrijednosti u klasi 1 i ($n + o$) vrijednosti u klasi 0) a, pritom, zavisna eksperimentalna varijabla Y (na kojoj je model

ugađan/optimiran) ima $(p + u)$ vrijednosti klase 1, i $(n + o)$ vrijednosti klase 0. Prema tome, izraz (1.2) temelji se samo na frekvencijama pojavljivanja klase 1 u eksperimentalnoj $(p + u)$ i predviđenoj varijabli $(p + o)$, i na frekvenciji pojavljivanja klase 0 u eksperimentalnoj $(n + o)$ i predviđenoj varijabli $(n + u)$.

Pretpostavljeno je da će provedbom istraživanja u disertaciji biti moguće simulacijama dobiti i vrijednosti za maksimalnu i minimalnu točnost (korelaciju), ali i izvesti izraze za izračun maksimalne i minimalne točnosti nasumičnoga modela. Minimalnu i maksimalnu točnost te najvjerojatniju nasumičnu točnost (kako je izraz (1.2) nazvan ranije [34,35] - a što će u disertaciji biti korigirano u „prosječna nasumična točnost“) zajedno nazivamo karakterističnim vrijednostima parametra točnosti $Q_2 = (p + n)/N$. Nadalje, tako definirane karakteristične vrijednosti parametra točnosti klasifikacijskog modela s dva stanja (Q_2) moguće je izračunati i za bilo koji parametar koji je mjera kvalitete modela. Analogno tome, pretpostavljeno je da je moguće izračunati i broj mogućih nasumičnih realizacija klasifikacijskih varijabli u ovisnosti o udjelima pojedinih klasa i odrediti informacijski sadržaj u njima.

Nadalje, pretpostavljeno je kako će na temelju tih izraza biti moguće izračunati parametre složenosti klasifikacijskih varijabli. Ti će se parametri složenosti primijeniti u analizi kvalitete klasifikacijskih varijabli kao deskriptora i u postupku izbora QSAR modela – što je glavna svrha istraživanja planiranih u disertaciji. Izraz (1.2) opisuje nasumičnu točnost klasifikacijskog modela s dva stanja koji predviđa $(p + o)$ vrijednosti u klasi 1 i $(n + o)$ vrijednosti u klasi 0), a pritom, zavisna eksperimentalna varijabla Y (na kojoj je model ugađan/optimiran) ima $(p + u)$ vrijednosti klase 1, i $(n + o)$ vrijednosti klase 0. Za uravnoteženi model koji predviđa jednaki omjer klasa kakav je i stvarni (eksperimentalni) omjer klasa u varijabli Y, izraz (1.2) za uravnoteženi model postaje (1.3)

$$Q_{2,rnd} = Q_{2,rnd-bal} = \frac{p + u\ ^2 +\ n + o\ ^2}{N^2}. \tag{1.3}$$

Ovaj izraz može se posve analogno primijeniti na analizu permutacijskog poklapanja varijable sa samom sobom. Naime, zamislimo klasifikacijsku varijablu s dvije vrijednosti (klase), tako da postoji samo jedna vrijednost u klasi 1, i devet vrijednosti u klasi 0 $(N = 10)$. Uvrste li se te vrijednosti u jednadžbu (1.3) dobijemo $Q_{2,rnd} = 0.1^2 + 0.9^2 = 0.82 = 82\ \%$. Broj mogućih poredaka (permutacija vrijednosti) te varijable daleko je manji nego je to kod varijable koja ima po pet vrijednosti u svakoj od dviju klasa [34], za koju vrijedi:$Q_{2,rnd} = 0.5^2 + 0.5^2 = 0.5 = 50\ \%$. Ta će se međuovisnost podrobnije istražiti u disertaciji, kao i sama vrsta funkcionalne veze između nasumične točnosti i broja mogućih permutacija modelne varijable. Očekuje se da će dobiveni rezultati te analize biti korisni u definiranju složenosti varijable.

Cilj istraživanja u sklopu izrade doktorske disertacije razdvaja se u tri pod-skupine aktivnosti koje se nadopunjuju. Prvi dio istraživanja vezan je uz definiranje i razvoj novih parametara složenosti klasifikacijskih varijabli (izvedenih/izračunanih iz strukture molekula). Taj dio istraživanja provodit će se teorijskim razmatranjem, istraživanjem i izračunom točnosti nasumičnog modela uporabom koncepata elementarne statističke analize i teorije vjerojatnosti. Definirat će se i izvesti izrazi za – najvjerojatniju, najmanju i najveću nasumičnu točnost uravnoteženih modela te će se u izravnoj analogiji primijeniti na izračun karakterističnih vrijednosti klasifikacijskih varijabli. Dobivene vrijednosti usporedit će se s odgovarajućim vrijednostima koji se dobivaju simulacijama za uravnotežene klasifikacijske modele postupkom koji se često koristi u literaturi, i sastavni je dio skupa pravila za provjeru kvalitete QSAR modela [16]. Taj se postupak (permutacije) provodi na

način da se jedna varijabla (obično Y) drži u stvarnom poretku, dok se druga (X, nezavisna) varijabla preslaguje u nasumičnim poretcima. Između Y varijable i svake permutirane varijable (kojoj su vrijednosti nasumično presložene) računat će se nasumična korelacija. Nakon određenog broja ponavljanja pokusa (obično manje od 1 ili 2 % svih mogućih permutacija varijable X) analizirat će se maksimalna postignuta korelacija. Ta će vrijednost služiti za ocjenu koliko je stvarna vrijednost koeficijenta korelacije varijabli Y i X (kad su one u stvarnim poretcima) veća od najveće vrijednosti dobivene (nasumične) simulacijama. Taj pokus analogno se može primijeniti na jednu varijablu, pri čemu se kao varijabla Y uzimaju vrijednosti varijable u stvarnom poretku, dok je modelna varijabla X je svaki put nasumično permutirana varijabla Y. Razlika između najmanje i najveće korelacije/točnosti dovest će se u vezu s entropijom (tj. informacijskim sadržajem) i složenošću varijable. Ti će izrazi (formule) u obliku novih parametara i postupaka za procjenu složenosti klasifikacijskih varijabli koristiti u daljnjoj provedbi istraživanja u disertaciji - i primijeniti u analizi složenosti klasifikacijskih varijabli.

Drugi dio istraživanja odnosit će se na komparativnu analizu izvedenih izraza i onih dobivenih simulacijama, te procjenu složenosti i razine nasumične korelacije te broja mogućih realizacija postojećih modela iz literature, kao i varijabli u njima (nezavisnih i zavisnih), kao i novih modela odnosa strukture i svojstava ili biološke aktivnosti molekula. U tu svrhu, koristit će se besplatno dostupne baze struktura i svojstava/aktivnosti bioaktivnih kemijskih spojeva i proteina.

Treći dio istraživanja odnosit će se na dizajniranje algoritma i računalnoga programa za izradu besplatno dostupnoga mrežnoga poslužitelja. Aplikacija (mrežni poslužitelj) je osmišljena da će preko sučelja korisnici moći učitati svoje datoteke s podacima (varijablama/deskriptorima) na server i provesti analize složenosti varijabli.

1.5. Očekivani znanstveni doprinos predloženog istraživanja

Parametri složenosti klasifikacijskih varijabli originalni su doprinos procjeni informacijskog sadržaja u varijablama koje su izračunane na temelju strukture molekula. Njihova uporaba bit će u analizi kvalitete varijabli u modelima odnosa između strukture i svojstava molekula (QSAR modeli), procjeni razine nasumične korelacije te u procjeni kvalitete samih modela.

Uspostavit će se odnosi između izvedenih i simulacijskih karakterističnih (najvećih, najmanjih i najvjerojatnijih/prosječnih) vrijednosti dobivenih permutacijskom analizom klasifikacijskih varijabli. To će doprinijeti boljem razumijevanju pozadine permutacijskih analiza modela definiranih OECD-ovim pravilima za regulatorne QSAR modele važnih za zaštitu ljudskoga zdravlja i okoliša [16].

Simulacije i teorijski razvoj u obliku izvedenih matematičkih izraza za izračun: (1) karakterističnih vrijednosti parametara za iskazivanje kvalitete modela, (2) entropiju varijable i (3) normaliziranu entropiju varijable, iskoristit će se za procjenu složenosti klasifikacijskih varijabli u QSAR modelima što je glavna svrha istraživanja u disertaciji. Rezultati dobiveni analizom složenosti klasifikacijskih varijabli s dva stanja moći će se primijeniti i u približnoj procjeni složenosti kontinuiranih varijabli, nakon što se provede njihova digitalizacija uporabom srednje vrijednosti varijable. Dobiveni rezultati bit će korisni za unapređenje postupaka (pre)selekcije varijabli u QSAR modelima koji se primjenjuju u brojnim područjima istraživanja u kemiji, bioznanostima i biotehnologiji, dizajniranja novih lijekova, istraživanja negativnog utjecaja kemijskih spojeva na okoliš i žive organizme (toksičnost, kancerogenost) do analiza i predviđanja

fizikalno-kemijskih svojstava kemijskih spojeva. Besplatnim mrežnim poslužiteljem razvijenim ovim doktorskim radom bit će omogućeno provjeriti kvalitetu varijabli uključenih u QSAR modele.

2. MATERIJALI I METODE

Prvi dio metoda odnosi se na simulacijska istraživanja, tj. na permutacijsku analizu klasifikacijskih varijabli korištenu u određivanju minimalnih, maksimalnih i prosječnih nasumičnih vrijednosti parametara (mjera) kvalitete modela prilagođenih analizi jedne varijable. Minimalnu, maksimalnu i prosječnu nasumična vrijednost zajednički nazivamo karakterističnim vrijednostima parametra kvalitete. Simulacije su provedene uz pomoć paketa R, u kojem su izrađeni potrebni postupci i računalni kod. Drugi dio rezultata dobiven je čisto teorijskim razmatranjima i izvođenjem karakterističnih vrijednosti parametara i entropije varijable, a pritom korištene metode temelje se na teorijskom znanju iz osnovama matematičke analize, teorije vjerojatnosti i algebre. Treći dio metoda korištenih u ostvarenju planiranih rezultata odnosi se na izradu mrežnog poslužitelja (servera). U podlozi tog dijela je programski kod izrađen tijekom provedbe istraživanja u disertaciji, a realiziran je u raznim programskim alatima. Naposljetku, za ilustraciju primjene dobivenih rezultata za procjenu složenosti varijabli opisani su skupovi podataka iz literature kao i oni kreirani u sklopu istraživanja u disertaciji. U prikazu i ilustraciji rezultata osim besplatnih programskih alata i paketa R korišteni su i komercijalni programi za obradu podataka i vizualizaciju.

2.1. Definicija pojmova i teorija

U istraživanjima u disertaciji uvedeni su novi matematički i informatički koncepti, postupci i pojmovi koje je potrebno najprije opisati, što je učinjeno u prvom dijelu ovog poglavlja. Ti koncepti osmišljeni su tijekom rada na istraživanjima u disertaciji, i nisu preuzeti iz literature, te jednim dijelom predstavljaju i metodološku novost koja se može smatrati i rezultatima istraživanja u disertaciji. Ti novi koncepti i postupci bili su nužni kako bi se mogli ostvariti planirani ciljevi istraživanja, i dobiti rezultati koji su opisani u narednom poglavlju disertacije.

2.1.1. Eksperimentalna i modelna klasifikacijska varijabla s dvije klase

Pojam varijable u ovom radu odnosi se na eksperimentalnu (E) i modelnu (M) klasifikacijsku varijablu s dva stanja. U eksperimentalnoj varijabli vrijednosti klasa (klasa 1 i klasa 0) odgovaraju onima iz eksperimenta kojim je određeno je li svojstvo ili aktivnost svake od molekula iz skupa koji se istražuje pripada klasi 1 ili klasi 0. U modelnoj varijabli klasa u koju pripada svojstvo ili aktivnost pojedine molekule iz skupa koji se istražuje dobivena je s pomoću modela. Važno je napomenuti da u istraživanjima u disertaciji modelna varijabla M ima (a) isti broj elemenata u klasi 1 i u klasi 0 kao i eksperimentalna varijabla E i (b) s obzirom da model nije savršen i da predviđa klasu molekule s određenom greškom. Naime, osim za sasvim točno predviđanje, oznake klase u varijablama E i M neće se podudarati za svaku molekulu, tj. neće se oznake klase 1 i oznake klase 0 nalaziti na istom mjestu u poretku 1, …, i, …, N u skupu molekula. Ali, kako je spomenuto pod (a), ukupni brojevi molekula čije svojstvo ili aktivnost pripadaju klasi 1 i klasi 0 bit će isti u varijablama E i M.

2.1.2. Tablica pogrešaka

U teorijskim i simulacijskim istraživanjima razmatraju se svojstva raznih statističkih parametara, a za njihov izračun potrebne su samo vrijednosti veličina p, n, u i o. Njihova značenja opisana su u nastavku.

Pozitivno ispravno predviđanje - p (*engl.* positive correct predictions) kada se vrijednost 1 (klasa 1) u obje varijable (E i M) nalaze se na istom mjestu. Kada je u E i M varijabli vrijednost 1, tada je molekula aktivna na tom mjestu i to je ispravno predviđeno. Ukupan broj takvih slučajeva označava se s p (pozitivno). Taj naziv dolazi iz podjele na 'pozitivnu' (+) i 'negativnu' (-) klasu.

Negativno neispravno predviđanje - u (engl. under-predictions) zbroj je svih slučajeva kada je u eksperimentalnoj varijabli (E) vrijednost jednaka 1, a u modeliranoj (M) vrijednost je jednaka 0

Negativno ispravno predviđanje - n (*engl.* negative correct predictions) je zbroj svih negativno ispravnih predviđanja kada se vrijednosti 0 u varijablama E i M nalaze na istom mjestu

Pozitivno neispravno predviđanje - o (*engl.* over-predictions) zbroj svih slučajeva u varijablama E i M kada je na istom mjestu u eksperimentalnoj varijabli (E) vrijednost 0, a u modeliranoj varijabli (M) vrijednost 1.

Zbroj svih parametara odgovara veličini svake od varijabli i njegova je vrijednost u većini simulacija u disertaciji jednaka 100.

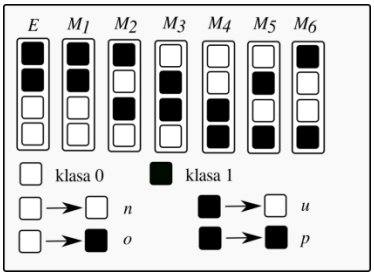
$$p + u + n + o = N = 100 \tag{2.1}$$

Značenje veličina p, n, u i o za slučaj varijabli E i M s dvije klase (1 i 0) ukratko je prikazano u *Tablici 2.1*, koja se računa i daje u svim istraživanjima i modeliranjima temeljenim na klasifikacijskim varijablama. Osim primjera s dvije klase u *Tablici 2.1*, analogno je moguće proširiti tablicu pogrešaka na slučaj klasifikacije s tri ili više klasa.

Tablica 2.1 Definicija tablice pogrešaka

		(predikcija) 1	(predikcija) 0	Σ redaka (eksperiment)
(eksperiment)	*1*	p	u	$p + u$
(eksperiment)	*0*	o	n	$n + o$
Σ kolona (predviđenih)		$p + o$	$n + u$	

Značenja veličina p, n, u i o, i njihova ovisnost o svim mogućim permutacijama modelne varijable M, grafički su ilustrirana na *Slici 2.1*. Na slici je crnom bojom označena klasa 1, a bijelom klasa 0. Varijable na *Slici 2.1* imaju $N = 4$ vrijednosti.



Slika 2.1 Grafički prikaz eksperimentalne (E) i modelnih varijabli (M_1, ... , M_6) dobivenih permutacijom eksperimentalne varijable i prikaz definicije elemenata tablice pogrešaka

Pri računanju vrijednosti p, n, u i o prva je varijabla uvijek eksperimentalna (E), a druga je jedna od modelnih (M_1, ..., M_6). Modelne varijable odnose se na varijable dobivene modeliranjem koje je u ovom slučaju modelnih varijabli M_i na *Slici 2.1* specifično te se provodi permutiranjem vrijednosti varijable E u svim mogućim različitim poretcima. Odgovarajuće vrijednosti p, n, u i o, parametri stvarne (Q_2) i nasumične točnosti $(Q_{2,rnd})$ te razlike stvarne i nasumične točnosti (ΔQ_2) dane su u *Tablici 2.2*. za sve moguće permutacije modelne varijable prema *Slici 2.1*.

Tablica 2.2 Elementi tablice pogrešaka na primjeru *Slike 2.1*

	Parametar						
Model	p	n	u	o	Q_2	$Q_{2,rnd}$	ΔQ_2
M_1	2	2	0	0	1.0	0.5	0.5
M_2	1	1	1	1	0.5	0.5	0
M_3	1	1	1	1	0.5	0.5	0
M_4	0	0	2	2	0	0.5	-0.5
M_5	1	1	1	1	0.5	0.5	0
M_6	1	1	1	1	0.5	0.5	0

Model 1 (modelna varijabla M_1) i model 4 (modelna varijabla M_4) izdvajaju se po vrijednostima p, n, o i u od ostalih slučajeva u *Tablici 2.2*. Modelna varijabla M_1 potpuno se poklapa u svim vrijednostima s eksperimentalnom varijablom E, a u slučaju modelne varijable M_4 riječ je o potpunom nepreklapanju s vrijednostima eksperimentalne varijable E.

Važno je ovdje naglasiti da se u literaturi spomenuti elementi matrice pogrešaka u području modeliranja sekundarne strukture proteina označavaju s p, n, o i u. Takav način označavanja preuzet je kao jednostavnija (kraćeg zapisa), ali i stoga što je korištena u prethodnim radovima hrvatskih autora [34,35]. U velikoj većini znanstvene i stručne literature iz područja medicinskih, informatičkih, računalnih i društvenih znanosti, elementi matrice pogrešaka označavaju se na slijedeći način:

- TP – (engl. *true positive*) = p
- TN – (engl. *true negative*) = n
- FN – (engl. *false negative*) = u
- FP – (engl. *false positive*) = o.

Imajući u vidu da je taj način označavanja elemenata tablice pogrešaka prevladavajući, konačni izvedeni izrazi u disertaciji mogu se jednoznačno i jednostavno prevesti na taj način označavanja uvođenjem supstitucija $p = TP$, $n = TN$, $u = FN$, i $o = FP$.

2.1.3. Parametar točnosti modela prilagođen analizi informacijskog sadržaja varijabli

Najjednostavniji i najčešće korišten parametar za izračun točnosti modela i za izračun podudarnosti dviju klasifikacijskih varijabli u primjerima sa *Slike 2.1* i *Tablice 2.2* naziva se parametar točnosti, ili postotne točnosti (jednadžba (2.2)):

$$Q_2 = \frac{(p+n)}{N} = \frac{100(p+n)}{N}\,(\%) \tag{2.2}$$

a njegove vrijednosti dane su za svih šest modelnih varijabli u *Tablici* 2.2. Vidimo da su vrijednosti tog parametra u rasponu od 0 do 1.0 ili od 0 % do 100 %. Dakle, Q_2 je omjer zbroja točnih predviđanja klase 1 i klase 0, i ukupnog broja svih elemenata klase 1 i klase 0 (N) u varijabli E odnosno M.

Međutim, poznato je da se ovisno o omjeru broja elemenata u varijabli koji pripadaju klasi 1 i klasi 0 – i nasumičnim pogađanjem mogu dobiti visoke vrijednosti parametra Q_2 [34]. Ono što nas zanima kod konkretnog modela (kada se želi izračunati njegova točnost predviđanja) njegov je doprinos (ΔQ_2) parametru točnosti Q_2 (jednadžba (2.3)):

$$\Delta Q_2 = Q_2 - Q_{2,rnd} \tag{2.3}$$

koji je iznad najvjerojatnije (ili prosječne) razine nasumične točnosti (podudarnosti) $Q_{2,rnd}$, koja se može dobiti nasumičnim pogađanjem vrijednosti 0 i 1. Nasumična točnost računa se iz elemenata tablice pogrešaka (*Tablica 2.1*) prema jednadžbi (2.4):

$$Q_{2,rnd} = 100\,\frac{(p+u)(p+o)+(n+o)(n+u)}{N^2}(\%) \tag{2.4}$$

Za modelne varijable sa *Slike 2.1* izračunane su u Tablici 2.2 vrijednosti $Q_{2,rnd}$ uvijek su istog iznosa i jednake su 0.5 (50 %), a vrijednosti parametra ΔQ_2 su u rasponu od -0.5 do 0.5. Razlog da je $Q_{2,rnd}$ = 0.5 u svim primjerima iz *Tablice 2.2* (i *Slike 2.1*) objašnjen je u Poglavlju 1.4, a može se protumačiti činjenicom da nasumična točnost prema jednadžbi (2.4) ovisi samo o udjelu klase 1 u varijablama E i M. S obzirom na to da je udio klase 1 određen, tim je automatski određen i udio klase 0 jer je riječ o dvoklasnim varijablama. U varijablama sa *Slike 2.1* i *Tablice 2.2* uvijek je podjednak broj elemenata klase 1 i klase 0 ($x = 1/2$,tj. udio svake klase je po 50 %) u varijablama E i M. U jedn. (2.4) broj elemenata klase 1 u varijabli E je $p+u$, a $n+o$ je broj elemenata klase 0 u varijabli E. Analogno je $p+o$, odnosno $n+u$ broj elemenata klase 1 odnosno klase 0 u varijabli M. Doprinos broja elemenata klase 1 u jednadžbi (2.4) je $p+u$ $p+o$ $= 2 \cdot 2 = 4$, a doprinos elemenata klase 0 jednak je $n+o$ $n+u$ $= 2 \cdot 2 = 4$. Kako je N = 4, dobiva se jednostavnim računom da je $Q_{2,rnd}$ = 0.5. Takav bi se rezultat dobio i ako bi samo jedna od varijabli E ili M imala 50 % udjela klase 1 (ili klase 0). Naime, u tom slučaju jednadžba (2.4) može se faktorizirati jer je $p+u = n+o = N/2$, dok preostali faktori zajedno u zbroju daju N. Jednostavnom pokratom N u brojniku i nazivniku, kao rezultat dobije se $Q_{2,rnd}$ = 0.5.

Početna ideja određivanja složenosti varijabli u ovom radu prije svega vezana je uz parametar ΔQ_2, koji računa točnost predviđanja (Q_2) vrijednosti varijable s dva stanja iznad nasumičnog predviđanja [34]. Parametar ΔQ_2 uveden je i uporabljen u analizi točnosti modela za predviđanje sekundarne strukture proteina [34]. Za njegov izračun potrebno je prethodno izračunati i najvjerojatniju nasumičnu točnost $Q_{2,rnd}$ prema jednadžbi (2.4). Ta je veličina početno definirana kao vrijednost parametra Q_2 koja se najčešće pojavljuje slučajnim izborom vrijednosti varijable [9]. Međutim, to vrijedi samo ukoliko je riječ o parnom broju podataka. Provedbom simulacija u disertaciji, pri čemu su permutacijom varijable E u svim poretcima kreirane sve moguće modelne varijable M (i pri čemu su se svaki put računale vrijednosti Q_2), dobiven je važan rezultat koji pokazuje da $Q_{2,rnd}$ iz [34] (formula (2.4)) odgovara prosječnoj nasumičnoj vrijednosti svih tako izračunanih vrijednosti Q_2. Uočava se to i na primjeru sa *Slike 2.1* gdje se vidi da je prosječna vrijednost svih parametra točnosti Q_2 iz *Tablice 2.2* jednaka 0.5, što je jednako vrijednosti parametra $Q_{2,rnd}$.

2.1.4. Ostali parametri kvalitete modela prilagođeni analizi informacijskog sadržaja varijabli

U disertaciji će biti prikazani rezultati za odabrane parametre čiji su izrazi dani formulama (2.5) – (2.9) koji će biti dobiveni odgovarajućim izvodima.

Pored parametra točnosti (Q_2) najčešće korišten parametar za procjenu točnosti (kvalitete) klasifikacijskih modela je Matthews-ov koeficijent korelacije (*MCC, engl.* Matthews Correlation Coefficient) dan formulom (2.5).

$$MCC = \frac{np - ou}{\overline{(p+o)(p+u)(n+o)(n+u)}} \tag{2.5}$$

MCC je u biti Pearson-ov koeficijent korelacije prilagođen radu s binarnim klasifikacijskim varijablama (tj. varijablama s dva stanja), a rad u kojem je izveden, iznimno je često citiran u znanstvenoj literaturi [36]. Prosječna apsolutna pogreška MAE (jednadžba u slučaju klasifikacijskih varijabli s dva stanja predstavlja kvadriranu standardnu pogrešku, tj. $MAE = s^2$ prema jednadžbi (2.6).

$$MAE = \frac{o+u}{N} \tag{2.6}$$

Standardna pogreška s (formula (2.7)) predstavlja drugi korijen omjera zbroja svih netočnih predviđanja u brojniku, i broja elemenata u varijabli u nazivniku (N) [36]:

$$s = \overline{\frac{o+u}{N}} \tag{2.7}$$

$F1 = F_1score$ (formula (2.8)) predstavlja omjer dvostrukog broja točno predviđanja manjinske klase ($2p$) i ($2p + u + o$) [37].

$$F_1score = F1 = \frac{2p}{2p+o+u} \tag{2.8}$$

Pritom je u formuli (2.8) ukupni broj točnih predviđanja većinske klase n (negativne klase koja je označena kao 0) zamijenjen (tj. izjednačen) s ukupnim brojem točnih predviđanja klase 1. To je specifična mjera kvalitete koja se koristi za procjenu točnosti modela pri predviđanju klasifikacijske varijable s dvije klase. Koristi se u slučajevima kada broj točno predviđenih elemenata većinske klase (n) nije poznat, ali je poznato da je taj broj neizmjeran, tj. puno veći od p, o i u, pojedinačno i zbirno. U disertaciji je razmatran i analiziran u analizi složenosti varijabli i parametar Cohenov kapa (κ) koji je definiran izrazom (2.9) kao omjer stvarnog doprinosa modela (ΔQ_2) i ($1 - Q_{2,rnd}$) [38].

$$\kappa = \frac{\Delta Q_2}{1 - Q_{2,rnd}} \tag{2.9}$$

Kako bi bilo moguće izračunati spomenute parametre, iz skupova je potrebno odrediti slijedeće veličine prikazane formulama (2.5) do (2.9):

Svaki od parametara u formulama (2.5) do (2.9) pojednostavljuje se u slučaju uravnoteženog modela (predviđanja) kada je $u = o$. Kako bi se jednostavnije objasnio taj postupak, potrebno je ući dublje u problematiku izmjenjivih varijabli.

2.1.5. Izmjenjive varijable

Izmjenjive varijable su one varijable koje imaju identičan broj elemenata u klasi 1 i elemenata u klasi 0. Permutirana varijabla M početne varijable E i sama početna varijabla E čine par izmjenjivih varijabli [39,40]. Važnost izmjenjivih varijabli je u tome da je pomoću njih moguće objasniti složenost varijabli na način da se promatra koliko je moguće napraviti različitih permutacija neke binarne varijable. Pojam izmjenjivih varijabli pronašli smo u matematičkoj literaturi u samoj završne faze pisanja disertacije. Taj pojam i koncepti koriste se u području obrade informacija s primjenom u računarstvu.

U svrhu pojednostavljenja izraza, uvodi se nova varijabla x koja je omjer ukupnog broja elemenata klase 1 u varijabli podijeljen ukupnim brojem elemenata u varijabli (N). Kako su u istraživanjima za oznake klasa korištene znamenke 0 i 1, tako je udio klase 1 (x) zapravo srednja vrijednost zbroja svih vrijednosti varijable E ili, analogno, modelne varijable (M). Označimo li eksperimentalnu varijablu kao $E = E(i)$ i modelnu varijablu kao $M = M(i)$ udio klase 1 računa se prema formuli (2.10))

$$x = \frac{\sum_{i=1}^{N} E(i)}{N} = \frac{\sum_{i=1}^{N} M(i)}{N} \tag{2.10}$$

Stvarni broj elemenata klase 1 u varijablama E ili M jednak xN ili X (formula (2.11)):

$$xN = X = p + u = p + o \tag{2.11}$$

Kod izmjenjivih varijabli, kakve su varijable E i M u disertaciji, broj netočnih negativnih predviđanja o uvijek je jednak broju netočnih pozitivnih predviđanja u ($o = u$) (*Prilog 2.1*), iz čega proizlazi $x_E = x_M$, tj. udjeli klase 1 jednaki su u varijablama E i M. Iz toga i iz jedn. (2.1) dobiva se relacija $N = p + 2u + n$.

Uz $o = u$, moguće je pojednostaviti izraze dane jednadžbama (2.4) do (2.9), a ta su pojednostavljenja dana jednadžbama (2.12) do (2.17).

$$Q_{2,rnd} = \frac{(p+u)^2 + (n+u)^2}{N^2}\,(\%) \tag{2.12}$$

$$\Delta Q_2 = \frac{N(p+n) - (p+u)^2 - (n+u)^2}{N^2} \tag{2.13}$$

$$MCC = \kappa = \frac{np - u^2}{(p+u)(n+u)} \tag{2.14}$$

$$MAE = \frac{2u}{N} \tag{2.15}$$

$$s = \sqrt{2\frac{u}{N}} \tag{2.16}$$

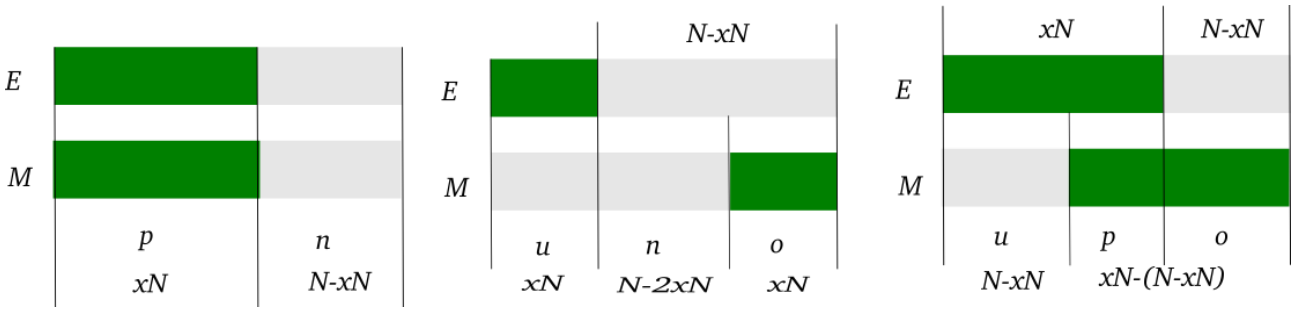
$$F1 = \frac{p}{p+u} \tag{2.17}$$

Za ove pojednostavljene izraze možemo reći da vrijede za izračun točnosti poklapanja para izmjenjivih varijabli, što je ekvivalentno paru (1) varijable E koja predstavlja eksperimentalne

vrijednosti i (2) odgovarajuće varijable M čiji su elementi dobiveni predviđanjem uravnoteženim modelom (za koji također vrijedi $o = u$).

2.1.6. Izvod izraza za elemente matrice pogreške u ovisnosti o udjelu klase 1 (x)

Koristeći kao promjenjivu varijablu udio klase 1 (x), moguće je izvesti izraze za minimalne i maksimalne karakteristične vrijednosti parametara kvalitete modela tako da se definiraju supstitucijski izrazi za veličine p, n, u i o koji se uvrštavaju u izvorne formule kojima se računaju ti parametri kvalitete. Kako bi se odredila funkcijska veza parametara p, n, u i o u ovisnosti o udjelu klase 1, napravljena je *Slika 2.2* gdje su označeni karakteristični dijelovi. Slika vrijedi samo za izmjenjive varijable – za slučajeve kada je jednak udio klase 1 u varijabli E i u varijabli M.



Slika 2.2 Pojednostavljeni prikaz različitih preslagivanja E i M varijable, te njihovih parametara i njihovih vrijednosti

Na *Slici 2.2* zelenom bojom označena su stanja 1, dok su sivom bojom označena stanja 0. Kako bi bilo moguće izvesti izraze za elemente matrice pogreške p, n, u i o u ovisnosti o x, koji će se koristiti u izvođenju minimalnih i maksimalnih karakterističnih vrijednosti parametara, potrebno je koristiti sortiranja vrijednosti varijabli E i M. Na lijevom dijelu *Slike 2.2.* prikazan je AA poredak varijabli E i M (obje varijable poredane jednako). Na srednjem i desnom dijelu iste slike prikazan je AD poredak kada su varijable E i M suprotno poredane – eksperimentalna varijabla uzlazno, modelna varijabla silazno. U srednjem dijelu *Slike 2.2* prikazan je slučaj kada je u varijabli manje od 50 % podataka klase 1, a na desnom dijelu *Slike 2.2* prikazan je slučaj kada je više od 50 % podataka klase 1. Ispod svakog segmenta prikazano je kojoj varijabli pripada i kako ga je moguće odrediti koristeći varijable x i N.

Svi izvedeni izrazi za elemente tablice pogrešaka p, n, u i o iskazane preko udjela klase i ukupnog broja podataka (N) za slučajeve varijabli E i M poredane na isti način (AA) i obrnuto (AD) dane su u *Tablici 2.3.*

Tablica 2.3. Supstitucijski izrazi za veličine p, n, u i o iskazane preko udjela klase 1 za izračun minimalne i maksimalne vrijednosti parametara (mjera kvalitete) klasifikacijskih varijabli s dva stanja

Varijabla	AD poredak	AA poredak
p	$0, \forall x \in [0, \frac{1}{2}]$	xN
	$(2x-1)N, \forall x \in [\frac{1}{2}, 1]$	
n	$(1-2x)N, \forall x \in [0, \frac{1}{2}]$	$N(1-x)$
	$0, \forall x \in [\frac{1}{2}, 1]$	
o, u	$xN, \forall x \in 0, \frac{1}{2}$	0
	$N\ 1-x\ , \forall x \in \frac{1}{2}, 1$	

U *Tablici 2.3* oznakom *AA* označeno je stanje jednako sortirane eksperimentalne (E) i modelirane (M) varijable (obje uzlazno), dok je oznakom *AD* označeno preslagivanje varijable E uzlazno, a varijable M silazno. Korištenjem izraza iz *Tablice 2.3*, moguće je odrediti minimalne, i maksimalne vrijednosti, i vrijednosti raspona parametara prikazanih formulama (2.12) do (2.17). Točnost formula danih u *Tablici 2.3* može se dokazati korištenjem slijedećih izraza: $xN = p + u = X,\ n + u = N(1-x)$ (*Prilog 2.3*). Već je poznato da je $o = u$ za slučaj izmjenjiviih varijabli E i M, a xN predstavlja ukupan broj elemenata (podataka) klase 1 u tim varijablama. Nadalje, $N(1-x)$ odnosi se na ukupni broj podataka (slučajeva) klase 0 u varijabli. Dakle, imamo ove identitete: *(Prilog 2.4 i 2.5)* $p + u = xN$ i $n + o = N(1-x)$ na cijelom području x (udjela klase 1).

2.2. Simulacijska istraživanja

2.2.1. Algoritam za provedbu simulacija

Simulacije su rađene na način da se kreiraju varijable s različitim omjerima klase 1 (x). I u prvom i u drugom slučaju riječ je o izmjenjivim varijablama, tj. varijablama *E* i *M* u kojima je omjer klasa 1 i 0 identičan. Simulacije su rađene na način da se dobije 100.000 varijabli *M* u raznim poretcima vrijednosti, odnosno dobije se isto toliko parova varijabli (*E* i *M*), pri čemu je varijabla *E* uvijek u istom (nepromijenjenom) poretku. U simulacijama za *N* = 100 i gdje je udio klase 1 (*x*) manji, pronaći će se i analizirati sve nasumične permutacije varijable *M*. Tada će minimalne, prosječne i maksimalne karakteristične vrijednosti parametara biti identične onima izvedenim formulama koje su apsolutno najmanje, najveće, odnosno identične prosječnim nasumičnim vrijednostima. To neće biti slučaj kod udjela klasa 1 i 0 između , npr. 70:30 % i 50:50 % gdje je broj mogućih permutacija jako velik. U tim slučajevima, simulacijama se uspijevalo dostići samo približnu najmanju i najveću te približnu srednju vrijednost svih izračunanih vrijednosti nekog parametra.

Svrha ovih simulacija je da se istraži ovisnost raspodjele vrijednosti različitih parametara kvalitete o omjerima klasa 1 i 0, te njihove rubne (najmanje i najveće) i prosječne vrijednosti.

Unutar pojedine simulacije, za svaki od 100.000 parova varijabli E i M računaju se u disertaciji vrijednosti ovih odabranih parametara: Q_2, $Q_{2,rnd}$, ΔQ_2, MAE, s, MCC, $F1$ i κ. Naposljetku, za svaki od tih parametara računaju se njegove najmanje i najveće te prosječne vrijednosti. Posebno su izdvojene dvije simulacije: u prvoj je omjer klasa u varijablama E i M jednak 50:50 %, a u drugoj omjer klasa u varijablama E i M iznosi 80:20 % (*Prilozi 3.42* i *E_3.5*).

Uporabom programskog jezika R napravljeno je ukupno 99 simulacija (*Prilozi 3.41* i *E_3.4*). Svaka simulacija uključuje eksperimentalne i modelirane varijable koje su binarni nizovi veličine 100 elemenata, te stupac koja sadrži informaciju o tome koliki je udio klase 1 u tom skupu. U ovom slučaju je riječ o udjelima varijabli od 1 do 99 % udjela klase 1 ($x \in [0,1]$). Za svaki par koji se sastoji od jednog eksperimentalnog niza i jednog modeliranog niza, određuju se elementi matrice pogrešaka p, n, u i o. Iz tih veličina, određuju se svi ostali parametri. Svrha simulacije je utvrditi približne raspone vrijednosti, a za manje udjele klase 1 i skoro posve točne rubne vrijednosti pojedinog parametra. Nadalje, karakteristične vrijednosti dobivene simulacijama poslužit će za provjeru točnosti karakterističnih vrijednosti izračunanih s pomoću izvedenih jednakosti. Usporedba ovisnosti izvedenih karakterističnih vrijednosti varijabli i njihovih raspona s entropijom bit će provedena pomoću korelacijskih analiza i prikaza ovisnosti.

Rezultati simulacija su grupirani i nad njima su vršene usporedbe s izvedenim podacima pomoću standardne pogreške te su izrađene vizualizacije. Proučavat će se povezanost s entropijom. U svrhu otkrivanja najprikladnijeg parametra za određivanje kompleksnosti varijabli. Povezanost varijabli s entropijama ispitivati će se metodom korelacije. Računat će se dvije vrste entropija; binarna entropija, te logaritmi ukupnog broja kombinacija.

2.2.2. Tehnologije u izradi simulacijskog algoritma

Već spomenute simulacije u poglavlju 2.2 koje služe kako bi se na velikom uzorku provjerila točnost formula ili omogućile regresijsku analizu tamo gdje formule nije lako dobiti. Prilikom izrade aplikacija za simuliranje i analizu podataka korišten je R jezik. U R izvornim kodovima korištene su slijedeće biblioteke:

combinat – stvaranje permutacija [41]

Compiler - ubrzanje rada funkcija kompajliranjem [42]

data.table - zamjena za data frame kod veće količine podataka [43]

ggplot2 - vizualizacija podataka [44]

sqldf - Biblioteka za korištenje SQL jezika nad podacima [4645]

scatterplot3d - 3D scatterplot vizualizacije [4746]

Metrics - Izračun RMSE [4847]

R.utils - Pomoć kod učitavanja podataka [4948]

R6 – biblioteka za rad s OOP (objektno orijentiranom paradigmom) u R-u [49]

readODS – biblioteka za čitanje i pisanje ODS datoteka [50]

2.3. Izvodi izraza za karakterističnih vrijednosti parametara i entropiju

Za parametar točnosti Q_2 prvobitno je planirano u disertaciji izvesti njegove karakteristične vrijednosti, tj. minimalne, maksimalne i prosječne vrijednosti. Također je planirano provesti simulacijsku analizu kako bi se i na taj način odredile karakteristične vrijednosti Q_2, te bi se tako provjerila i ispravnost relacija izvedenih teorijskim razmatranjima i analizama. Za potrebe izračuna minimalne, maksimalne i prosječne vrijednosti te raspona parametara $MCC, MAE, s, F1$ i κ korištene su simulacije, ali se pokazalo da ih je moguće dobiti i izvodima, što predstavlja važnu nadogradnju, proširenje i poopćenje očekivanih i planiranih rezultata istraživanja u disertaciji. Prilikom izvođenja parametara kvalitete korištene su formule iz *Tablice 2.3* (tablica transformacija) i uvrštavane u izvorne formule. Na taj način određuju se minimalne i maksimalne vrijednosti parametara, i njihovi rasponi.

Za izračun prosječnih nasumičnih vrijednosti parametara koristit će se vrijednosti iz posebne tablice supstitucija koja je napravljena za tu namjenu. Svi parametri bit će prikazani i u ovisnosti o parametru x koji se odnosi na udio klase 1 unutar odabrane varijable s N vrijednosti.

Za većinu parametara minimalne vrijednosti određuju se tako da se varijable poredaju suprotno, tako da je eksperimentalna varijabla uzlazno sortirana, a modelirana varijabla silazno (oznaka AD). Prilikom AD sortiranja svi parametri neće poprimiti minimalne vrijednosti nego će u slučajevima standardne pogreške s i srednje apsolutne pogreške MAE značiti suprotno.

Važno je uz to napomenuti da kod varijable s AD poretkom postoji lijeva i desna strana jednadžbe, tj. formule su definirane na rasponu lijevo i desno u odnosu na $x = 1\ \ 2$. Lijeva strana intervala označena s L je $\forall x \in\ 0,1\ \ 2\ \ $ a desna strana označena s R je $\forall x \in [1\ \ 2\ , 1]$. Razlika izraza za minimum i maksimum vrijednosti parametara varijabli dat će funkciju raspona varijable u ovisnosti o udjelu klase 1 ($x = (p + u)\ /\ N$).Za varijable koje nemaju nužno jednake udjele klasa, bit će uveden novi parametar $y =\ \ p + o\ \ \ N\ $ koji predstavlja udio klase 1 u varijabli M.

Osnovni izraz za entropiju koji odgovara permutacijskoj analizi varijable provedenoj u disertaciji odgovara fizikalnom konceptu entropije opisanom Boltzmanovom formulom $S =\ klnW$. Polazeći od tog izraza, teorijskim razmatranjima izveden je izraz za entropiju varijable u ovisnosti o udjelu klase 1 (x). Potom, taj je izraz normiran i korišten u procjeni složenosti varijable. Koncept entropije u analizi informacijskog sadržaja molekularnih deskriptora rijetko se pojavljuje u literaturi, a jedna od primjena koja koristi Shannonov koncept entropije dana je u radu [51].

2.4. Tehnologije korištene u izradi mrežnog poslužitelja i u analizama podataka

U ovom poglavlju opisane su aplikacije napravljene u svrhu analize podataka. Prilikom izrade mrežnog poslužitelja korišten je virtualni Linux server (virtualni PC) dobiven na korištenje od Sveučilišnog Računskog Centra (SRCE). Na serveru je instaliran Ubuntu Linux 18.04. i paket R verzija 3.60 [52].

2.4.1. Programske tehnologije u izradi aplikacije za izračun topoloških deskriptora

Web aplikacija „*Zagreb indices and their modifications – CALCULATOR*“ (opisana u *Prilogu 3.43*) napravljena je u svrhu pretvaranja MOL/SDF formata pretvorila u graf te omogućila računanje topoloških deskriptora.

Aplikacija je dostupna na adresi http://meteo2.irb.hr/indexer/ [53] i izrađena je kombinacijom više tehnologija. Dio koji računa valencije veza rađen je u tehnologiji Java [54]. Dodatne biblioteke korištene su pomoću Maven framework-a [55]. Sučelje je rađeno tehnologijom bootstrap koristeći pri tome jQuery library [56,57]. Ulazne datoteke moraju biti u MOL/SDF formatu [13], a rezultati se spremaju u CSV datoteke koje su dostupne za preuzimanje sa servera.

2.4.2. Programske tehnologije u izradi aplikacije simulatora

Programski kod za provedbu simulacija, naknadne analize i transformacije podataka rađen je u programskom jeziku R [58]. Za manipulacije podacima korišten je data.table library [43], a za vizualizaciju podataka ggplot2 [44] i plot3d [59]. Za filtriranje podataka korištena je sqlite baza u obliku sqldf biblioteke [46].

Vizualizacija rezultata simulacijskih analiza karakterističnih vrijednosti parametara kvalitete modela rađena je programom Origin [60], a dio proračuna rađen je i programom MS Excel [61].

2.4.3. Programske tehnologije u izradi aplikacije ProtSeqAnalizer

Računalna aplikacija ProtSeqAnalizer izrađena je u svrhu analize motiva unutar odabranih proteinskih sljedova. Ona omogućava grupiranje i filtriranje aminokiselina, te njihovo dodjeljivanje grupama.

U izradi korišten je RStudio [62]. Prilikom pisanja mrežne aplikacije korištene su data.table objekti za manipulaciju podacima [43]. Više informacija o aplikaciji dostupno je u elektroničkim prilozima (Prilozi - *Prilog 3.5*).

U analizama i primjeni bit će prikazan izračunani najjednostavniji motivi dviju susjednih aminokiselina, npr. alanin-valin ('AV') ili leucin-tirozin ('LT'), uzimajući pritom u obzir poredak (AV nije isto što i VA). Ti će motivi biti računani na skupu od 568 peptida [63].

2.4.4. Programske tehnologije u izradi mrežnog poslužitelja za analizu složenosti varijabli

Za primjenu rezultata, analize skupova deskriptora iz literature formulama izvedenim u disertaciji (Rezultati – formule 3.1 – 3.17), napravljena je mrežna aplikacija „*Classification variable complexity parameter estimator*“. Više informacija o programskom kodu aplikacije, dostupno je u elektroničkim prilozima. U njegovoj izradi korišten je RStudio [62], a web aplikacija rađena je tehnologijom R Shiny [62,64]. Prilikom pisanja web aplikacije korištene su data.table objekti za manipulaciju podacima [43].

2.5. Baze podataka za ilustraciju primjene rezultata i usporedbe

2.5.1. Baza antimikrobnih peptida DADP

Baza peptida DADP (*engl*. Database of Anuran Defense Peptides) preuzeta sa servera http://split.pmfst.hr/dadp [65], sadrži sekvence 568 antimikrobnih peptida žaba i eksperimentalne podatke o minimalnoj inhibitorskoj koncentraciji (MIC) za bakterije *E. coli* i *S. aureus* (MIC_ec, MIC_sa). Osim toga, baza sadrži i izmjerene vrijednosti 50 % hemolitičke aktivnosti (HC50) na stanicama eritrocita (koncentracija peptida pri kojima 50 % stanica eritrocita postane nefunkcionalno).

Aplikacijom ProtSeqAnalizer bit će istraživani strukturni motivi koji bi mogli biti korisni kao varijable i pridonijeti u modeliranju aktivnosti peptida (HC50, MIC_ec, MIC_sa) ili terapeutskog indeksa (TI) koji je definiran kao omjer HC50 i minimalne inhibitorne koncentracije (MIC_ec ili MIC_sa). [63]

2.5.2. Baze molekularnih deskriptora iz literature

U svrhu provjere zadovoljavaju li podaci iz drugih objavljenih radova kriterije složenosti izvedene u istraživanjima u sklopu izrade disertacije, napravljene su analize na skupovima deskriptora iz literature. To su skupovi topljivosti velikog broja organskih spojeva u vodi iz rada Huuskonen i drugi [66]. Također, izdvojeni su i priređeni te analizirani skupovi deskriptora iz rada [67] u kojem su razvijani multivarijatni QSAR modeli za predviđanje bioloških aktivnosti analoga taksana i pacitaksela.

2.5.3. Skup podataka za primjenu dobivenih rezultata u analizi kvalitete modela

Na priređenom skupu podataka prema ref. [35] primijenjeni će biti dobiveni rezultati i parametri kako bi se procijenila kvaliteta modela. Modeli su dobiveni različitim metodama u bioinformatici u sklopu natjecanja u predviđanju mutacija koje dovode do tumora. [35] Radi se o elementima matrice pogrešaka različitih modela na temelju kojih su oni rangirani u završnoj fazi (IS3) prediktivnog natjecanja za predviđanje tumorskih mutacija [35]. Priređena tablica dostupna je u dva dijela u *Prilozima 3.46 i 3.47*. Tablica je važna za usporedbu dosadašnjeg načina rangiranja modela s onim koji je rezultat preporuka proizašlih iz istraživanja provedenih u ovoj disertaciji.

3. REZULTATI

U ovom dijelu biti će prikazani rezultati simulacija, opis rada programa napravljenih u svrhu ovog rada i prikazani izvodi karakterističnih vrijednosti parametara.

3.1 Izvodi karakterističnih vrijednosti parametara Q_2 i ΔQ_2

U rezultatima najprije će biti izložen postupak izvođenja formula za računanje karakterističnih vrijednosti parametara, tj. njihove najmanje i najveće vrijednosti te njihove razlike pomoću udjela klase 1 (x). Potom će se pokazati da se vrijednosti parametara točnosti (Q_2), nasumične točnosti $Q_{2,rnd}$, i njihove razlike (ΔQ_2, nazvane u [34,35] stvarnim doprinosom modela iznad nasumičnog pogađanja - nasumične točnosti) mogu izraziti preko udjela klase 1, tj. $x = (p + u)/N$. Nadalje, pokazat će se da su razlike između najveće i najmanje vrijednosti, kao i između najveće i prosječne nasumične vrijednosti to manje što je udio većinske klase bliži 1. Intuitivno je jasno kako je u tom slučaju varijabla manje složena, i sadrži manju količinu informacije. Parametar Q_2 iskazuje točnost modela. Primijenjeno na analizu složenosti (ili informacijskog sadržaja) varijable, parametar Q_2 iskazuje podudarnost vrijednosti varijable E s odgovarajućim vrijednostima iste takve varijable permutirane u svim poretcima. Oznaka E predstavlja eksperimentalnu varijablu i ona ima stalni (fiksni) poredak vrijednosti. S M označava se modelna varijabla koja je zapravo varijabla E permutirana u raznim porecima u odnosu na početni stalni poredak vrijednosti varijable E. S obzirom da je razlika između E i M samo u rasporedu vrijednosti, jasno je da obje varijable imaju isti udio klase 1 (x), a time i udio druge klase koju označavamo s 0.

Parametar $Q_{2,rnd}$ prosječna je nasumična točnost modela uvedena i objašnjena u radu. [34] Osim prosječne nasumične točnosti, možemo zamisliti da se permutiranjem varijable M u odnosu na varijablu E može dobiti i lošije i bolje poklapanje vrijednosti tih dviju varijabli od prosječne točnosti iskazane $Q_{2,rnd}$. Točnost Q_2 može se izračunati za svaku permutaciju varijable M u odnosu na varijablu E.

Simulacijske analize u ovom radu pokazat će da je vrijednost $Q_{2,rnd}$ jednaka prosječnoj vrijednost parametra Q_2 izračunatoj između varijable E (svaki put u stalnom poretku) i njoj odgovarajuće modelne varijable M permutirane u svim poretcima u odnosu na varijablu E. Razlika parametara Q_2 i $Q_{2,rnd}$ daje parametar ΔQ_2, koji u općenitom slučaju predstavlja doprinos modela.

Minimalne i maksimalne karakteristične vrijednosti parametara dobivene su koristeći tablicu supstitucija za vrijednosti $p, n, u,$ i o iskazanih preko udjela klase 1 (x) u *Tablici 2.3*. Prosječne nasumične vrijednosti dobivene su uvrštavanjem supstitucijskih izraza za srednje vrijednosti $p, n, u,$ i o temeljene na teoriji vjerojatnosti [68].

Minimalne i maksimalne karakteristične vrijednosti parametara dobivene su koristeći tablicu supstitucija za vrijednosti p, n, u i o iskazanih preko udjela klase 1 (x) u *Tablici 2.3*. Prosječne nasumične karakteristične vrijednosti dobivene su uvrštavanjem supstitucijskih izraza za srednje vrijednosti elemenata matrice pogrešaka p, n, u i o temeljene na teoriji vjerojatnosti [68] i na analogiji s izračunom parametra nasumične točnosti $Q_{2,rnd}$ (jednadžba (1.2) i literatura [34,35]).

3.1.1 Izvodi karakterističnih vrijednosti Q_2 i ΔQ_2 u ovisnosti o udjelu klasa

Izvodi minimalnih i maksimalnih vrijednosti parametra Q_2 u ovisnosti o udjelu klase 1 dobiveni su uporabom supstitucijskih izraza iz Tablice 2.3 (*Materijali i metode*). Izraz (formula) za maksimalnu vrijednost parametra Q_2 izvodi se kada su varijable E i M poredane identično jedna u odnosu na drugu (poredak AA). Izraz za minimalnu vrijednost parametra Q_2 izvodi se kada su varijable E i M poredane (sortirane) suprotno jedna u odnosu na drugu (poredak varijabli AD) – varijable E i M poredane suprotno su poredane (sortirane) jedna u odnosu na drugu. Pri izvodu minimalne vrijednosti interval se razdvaja na dva pod-intervala vrijednosti x: (1) za lijevi pod-interval ($x \leq 1\ \ 2$) za koji se u indeksu pridodaje oznaka L (engl. *Left*), i (2) za desni pod-interval (x $\geq$ 1 2) za koji se u indeksu pridodaje oznaka R (engl. *Right*).

Minimalna vrijednost parametra Q_2 u ovisnosti o udjelu klase 1 dana je formulama (3.1) i (3.2), pri čemu su vrijednosti elemenata matrice pogrešaka $p, n, u,$ i o preuzete iz tablice supstitucijskih izraza za određivanje minimuma i maksimuma (*Materijali i metode – Tablica 2.3*). Izvodi formula (3.1) i (3.2) nalaze se *u Prilozima 3.1 i 3.2.*

$$Q_{2,AD,L} = 1 - 2x, \forall x \in [0, \frac{1}{2}] \tag{3.1}$$

$$Q_{2,AD,R} = 2x - 1, \forall x \in [\frac{1}{2}, 1] \tag{3.2}$$

Maksimalna vrijednost parametra Q_2 kao funkcija udjela klase 1 (x) prikazana je formulom (3.3) i vrijedi za cijeli interval $x \in [0,1]$

$$Q_{2,AA} = 1 \tag{3.3}$$

Iz minimalne ($Q_{2,min}$) i maksimalne ($Q_{2,max}$) vrijednosti, koje su dvije karakteristične vrijednosti parametra točnosti Q_2, moguće je izračunati raspon u kojem se javljaju sve vrijednosti Q_2. Apsolutni raspon vrijednosti parametra točnosti označavat ćemo zagradama ispred kojih dolazi znak Δ, tj. kao $\Delta(Q_2)$. To je potrebno kako bi se ta oznaka razlikovala od ranije uvedene oznake ΔQ_2 u radu Batiste i dr. [34] koja predstavlja razliku između maksimalne vrijednosti Q_2 i nasumične točnosti $Q_{2,rnd}$, kao treće karakteristične vrijednosti parametra Q_2.

Apsolutni raspon parametra Q_2 kao funkcija udjela klase 1 dobije se kao razlika izraza (3.3) i izraza (3.1) ili (3.2), i to posebno za x 1 2 i posebno za $x \geq$ 1 2. Na taj način dobivaju se izrazi (3.4) i (3.5) gdje se $\Delta Q_{2,L}$ odnosi na apsolutni raspon parametra točnosti na lijevom ($x \leq$ 1 2) a $\Delta Q_{2,R}$ na desnom ($x \geq$ 1 2) pod-intervalu udjela klase 1.

$$\Delta(Q_{2,L}) = Q_{2,max} - Q_{2,min} = Q_{2,AA} - Q_{2,AD,L} = 2x, \forall x \in [0, \frac{1}{2}] \tag{3.4}$$

$$\Delta(Q_{2,R}) = 2(1 - x), \forall x \in [\frac{1}{2}, 1] \tag{3.5}$$

Izvodi formula (3.4) i (3.5) nalaze se u *Prilozima 3.4 i 3.5*. Apsolutni raspon parametra Q_2 u ovisnosti o udjelu klase 1 može se sažetije izraziti formulom (3.6) za dva podsegmenta:

$$\Delta(Q_2) = \begin{matrix} 2x, \forall x \in [0, \frac{1}{2}] \\ 2(1-x), \forall x \in [\frac{1}{2}, 1] \end{matrix} \quad (3.6)$$

Ti se izrazi mogu objediniti i sažetije prikazati za cijeli segment ($x \in [0,1]$) uporabom apsolutnih vrijednosti (formule (3.7) i (3.8)). Izraz (3.8) identičan je izrazu (3.7), samo je iskazan u postotcima:

$$\Delta(Q_2) = 1 - |1 - 2x|, \forall x \in [0,1] \quad (3.7)$$

$$\Delta(Q_2) = 100 - \ 100 - 200x \ (\%), \forall x \in [0,1] \quad (3.8)$$

S druge strane, ΔQ_2 predstavlja raspon točnosti predviđanja modela Q_2 (ili podudarnosti izmjenjivih varijabli E i M u analizama u disertaciji) koji je iznad nasumične točnosti. Pored izvoda tog parametra temeljenog na teoriji vjerojatnosti u radu Batiste i dr. [34], prosječna nasumična točnost $Q_{2,rnd}$ može se dobiti tako da se umjesto udjela p/N ili n/N u izrazu za izračun parametra Q_2 uvrste supstitucije $((p+o)(p+u))/N$ za p, te $(n+o)(n+u)/N$ za n. Tako se dobije formula (3.9) [34] koja sadrži elemente tablice pogrešaka (*Materijali i metode Tablica 2.1*) i ukupni broj (N) elemenata klase 1 i 0 u varijabli E odnosno M.

$$Q_{2,rnd} = \frac{(p+o)(p+u) + (n+o)(n+u)}{N^2} \quad (3.9)$$

Za izmjenjive varijable imamo $u = o$, pa se izraz za $Q_{2,rnd}$ pojednostavljuje, što je prikazano formulom (3.10):

$$Q_{2,rnd} = (\frac{p+u}{N})^2 + (\frac{n+u}{N})^2 \quad (3.10)$$

Kako parametar $Q_{2,rnd}$ iz jedn. (3.10) ovisi samo o udjelu klase 1 (prvi član) i udjelu klase 0 (drugi član), iz toga proizlazi da njegova vrijednost za jednu definiranu frekvenciju klase 1 nije ovisna o promjeni rasporeda vrijednosti u varijablama E i M, tj. vrijednost nasumične točnosti nije ovisna o permutacijama varijabli.

Uvrštavanjem elemenata supstitucijske tablice (Materijali i metode, *Tablica 2.3*, *AD* preslagivanje) dobiva se minimalna vrijednost parametra $Q_{2,rnd}$ kao funkcija udjela klase 1 prikazana je formulom (3.11) za slučajeve kad je udio klase 1 $x \leq 1\ \ 2$. Cijeli izvod dostupan je u prilozima (*Prilog 3.6*). Minimalna i maksimalna vrijednosti parametra $Q_{2,rnd}$ izražene preko udjela klase 1 jednake su (formula (3.11))

$$Q_{2,rnd,AD,R} = Q_{2,rnd,AD,L} = Q_{2,rnd,AA} = 2x^2 - 2x + 1, \forall x \in [0,1] \quad (3.11)$$

pa je, stoga, i njihova razlika jednaka nuli (formula (3.12)):

$$\Delta Q_{2,rnd} = 0 \quad (3.12)$$

Rezultat iz jedn. (3.11) očekivan je, jer je primijećeno ranije (poglavlje 1.4) da nasumična točnost ovisi samo o udjelima klase u varijablama koje se uspoređuju, a ne o poretku (redoslijedu) vrijednosti varijabli. Kako su minimalne i maksimalne vrijednosti parametara u disertaciji izvedene originalnim postupcima u kojima se varijable preslaguju u suprotnim (AD) ili istim (AA) redoslijedima, uporabom supstitucijskih vrijednosti iz *Tablica 2.3*, ovaj (očekivani) rezultat u jedn.

(3.11) potvrđuje ispravnost originalnog postupka izvođenja minimalnih i maksimalnih vrijednosti parametra točnosti Q_2.

Karakteristične minimalne i maksimalne vrijednosti parametra ΔQ_2 (doprinosa modela iznad nasumične točnosti) u ovisnosti o udjelu klase 1, izvedene su postupkom korištenim u izvođenju jednadžbi (3.1) i (3.2) pri čemu su za elemente matrice pogrešaka $p, n, u,$ i o preuzete vrijednosti iz tablice supstitucijskih izraza (*Tablica 2.3*). Minimalna vrijednost ΔQ_2 u ovisnosti o udjelu klase 1 (x) dobiva se obrnutim (AD) preslagivanjem vrijednosti varijabli E i M za $x \leq 1\ \ 2$, što daje izraz (3.13):

$$\Delta Q_{2,AD,L} = -2x^2, \forall x \in [0, \frac{1}{2}] \tag{3.13}$$

Cijeli izvod dan je u *Prilogu 3.9*. Za udjele klase 1 u pod-intervalu $x \geq 1\ \ 2$, minimalna vrijednost parametra ΔQ_2 određuje se formulom (3.14) čiji se izvod nalazi u prilozima (*Prilog 3.10*).

$$\Delta Q_{2,AD,R} = -x(x-1)^2, \forall x \in [\frac{1}{2}, 1] \tag{3.14}$$

Maksimalna vrijednost parametra ΔQ_2, kao funkcija udjela klase 1 (x) za $x \in\ 0{,}1$, dobivena uparenim preslagivanjem AA, iskazana je formulom (3.15), a njen izvod dan je u *Prilogu 3.11*.

$$\Delta Q_{2,AA} = -2x(x-1), \forall x \in [0,1] \tag{3.15}$$

Raspon (oznaka: $\Delta(\Delta Q_2)$) jednak je razlici njegovih maksimalnih i minimalnih vrijednosti. Minimalna i maksimalna vrijednost parametra je ovisna o $Q_{2,rnd}$, koji ne ovisi o preslagivanjima varijabli. Iz tog razloga raspon parametra ΔQ_2 označen kao $\Delta(\Delta Q_2)$ jednak je rasponu parametra Q_2 koji se označava kao ΔQ_2, što je prikazano formulom (3.16).

$$\Delta(\Delta Q_2) = (Q_2 - Q_{2,rnd})_{AA} - (Q_2 - Q_{2,rnd})_{AD} = Q_{2,AA} - Q_{2,AD} = \Delta Q_2 \tag{3.16}$$

Prosječna nasumična vrijednost parametra ΔQ_2 jednaka je nuli.

$$(\Delta Q_2)_{rnd} = 0 \tag{3.17}$$

3.1.2 Simulacije karakterističnih vrijednosti parametara točnosti

U ovom dijelu bit će objašnjene simulacije parametara s različitim udjelima klase 1 u izmjenjivim varijablama E i M. Svrha tih simulacija prvobitno je bila odrediti karakteristične vrijednosti parametara, osobito prosječne nasumične vrijednosti, za koje nije bilo poznato mogu li se odrediti analitički (matematičkim analizama i algebarskim izvorima). Minimalne i maksimalne vrijednosti parametra točnosti ΔQ_2 planiralo se dobiti uparenim i obrnutim sortiranjem varijabli E i M. Pretpostavljalo se da će se prosječne/najvjerojatnije nasumične vrijednosti nalaziti na maksimumu raspodjele simuliranih vrijednosti. Dodatni cilj simulacija bio je istražiti ovisnost karakterističnih vrijednosti parametra točnosti Q_2 i ΔQ_2 o različitim omjerima klase 1 i 0.

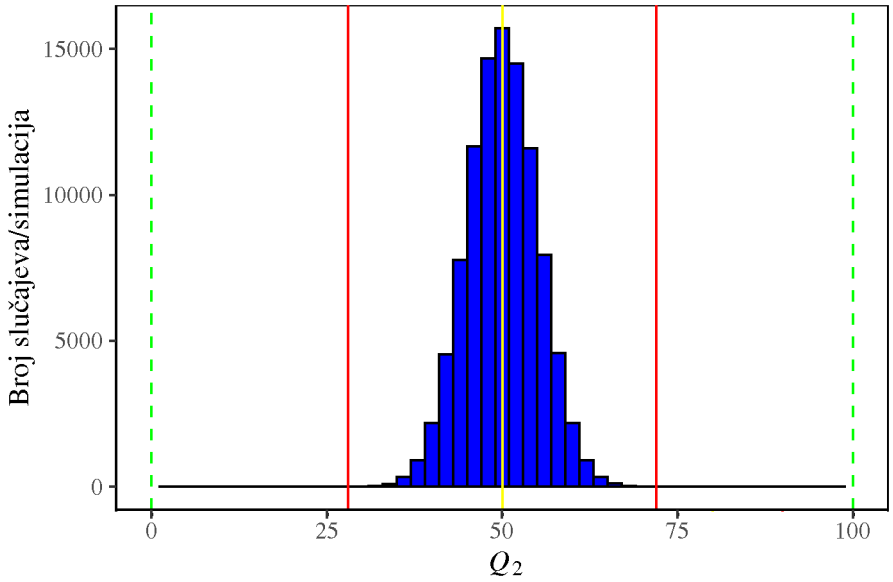
Među omjerima udjela klasa 1 i 0 izdvaja se omjer 50:50 %, pri kojem je broj mogućih permutacija varijable M najveći među svim omjerima. U drugom simulacijskom eksperimentu omjer klase 1 i 0 u varijablama E i M iznosi 80:20 %. Apsolutna minimalna i maksimalna vrijednost dobije se obrnutim i uparenim sortiranjem vrijednosti varijabli E i M, a pojedina

simulacijska vrijednost svakog parametra dobivena je na temelju izravnog izračuna parametara preklapanjem (usporedbom) varijabli E i M bez njihovog prethodnog sortiranja.

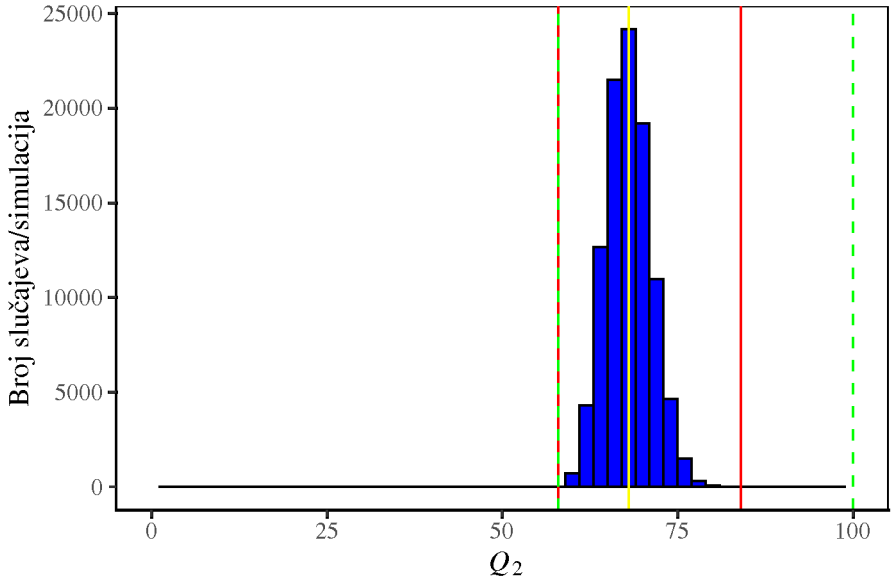
Simulacija se provodi na taj način da se omjer klasa 1 i 0 u varijablama E i M postavi na željenu vrijednost (50:50 % ili 80:20 %). Poredak vrijednosti u varijabli E drži se stalnim, dok se vrijednosti varijable M permutiraju nasumično u 100.000 poredaka, pri čemu se dobiva isto toliko parova izmjenjivih varijabli čija se podudarnost analizira u svakom koraku. Pritom je moguće da su se neke permutacije među 100.000 permutacija varijable M ponovile – jednom ili više puta. Obrada je izvršena aplikacijom razvijenom u okruženju R s pomoću skripti „test11.R" i „test12.R" koje su objašnjene u *Prilogu 3.42*.

Parametar točnosti – Q_2

Rezultati su spremljeni u „out" mapu u datoteke „test11.txt" (udio klase 1 je 50 %) i test12.txt (udio klase 1 je 80 %) (*Prilog 3.42*). Rezultati simulacija raspodjele vrijednosti dobivenih na temelju 100.000 nasumičnih parova varijabli E i M (pri čemu je varijabla E uvijek u istom poretku) prikazani su u obliku histograma za parametre Q_2 (*Slika 3.1*) i ΔQ_2. (*Slika 3.2*). Crvene vertikalne crte označavaju raspon vrijednosti dobiven simulacijama, a isprekidane vertikalne zelene crte apsolutno najmanji i najveći raspon vrijednosti parametra Q_2.



Slika 3.1 Raspodjela parametra kvalitete Q_2 za varijable E i M s omjerom klasa 50:50 %



Slika 3.2 Raspodjela parametra kvalitete Q_2 za varijable E i M s omjerom klasa 80:20 %

Razlike karakterističnih vrijednosti raspodjela parametra Q_2 prikazanih na *Slika 3.1 i 3.2* grupirani su i prikazani u *Tablici 3.1*.. Jasno je vidljiva razlika za različite udjele klase 1 i 0 između raspodjela – uočava se jasan pomak prema višim vrijednostima sredine raspodjela od omjera 50:50 % do omjera 80:20 %. Srednja vrijednost kod varijabli s omjerom klasa 80:20 % iznosi 68.01, a 50.02 kod varijable s omjerom klasa 50:50 %. Istovremeno, raspodjela je uža kod neravnotežnih udjela klasa 1 i 0, što znači da se smanjuje razlika između apsolutne minimalne i maksimalne vrijednosti od 100 % na 40 %. Također, raspodjela je asimetrična u eksperimentu kad je omjer klasa 80:20 % (ili općenito - kada omjer odstupa od omjera 50:50 %).

Tablica 3.1 Usporedba simulacijskih karakterističnih vrijednosti parametra Q_2 za udjele klase 1 50 % i 80 %

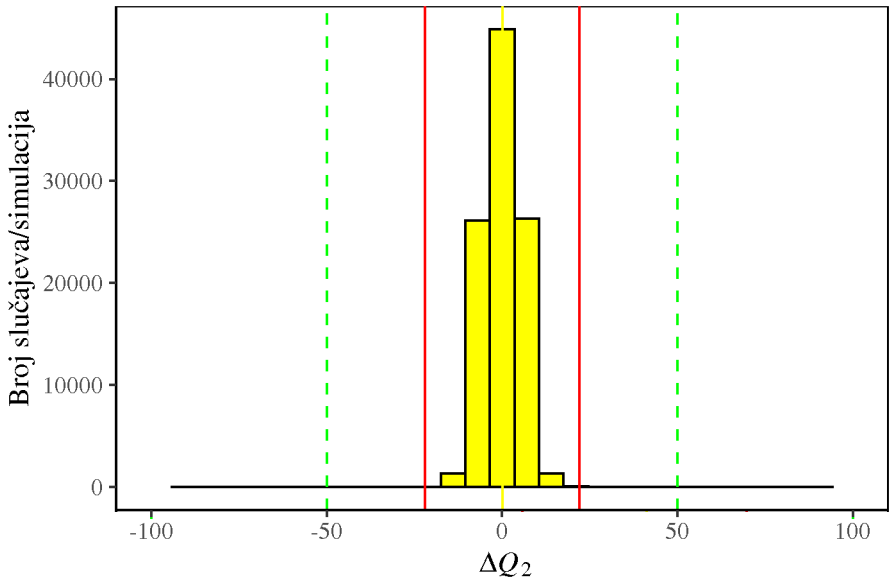
	$Q_{2,50}$	$Q_{2,80}$
Minimum	28.00	60.00
Maksimum	72.00	84.00
Medijan	50.00	68.00
Srednja vrijednost	50.02	68.01
Apsolutni minimum	0	60
Apsolutni maksimum	100	100

Raspon simulacijskih vrijednosti je $\Delta(Q_{2,S,50})$ = 44, odnosno $\Delta\ Q_{2,S,80}$ = 24, što ukazuje na smanjenje raspona parametra Q_2 povećanjem neuravnoteženosti udjela klasa, a slično je kod apsolutnih vrijednosti kod kojih je $\Delta\ Q_{2,A,50}$ = 100, dok je $\Delta\ Q_{2,A,80}$ = 40.

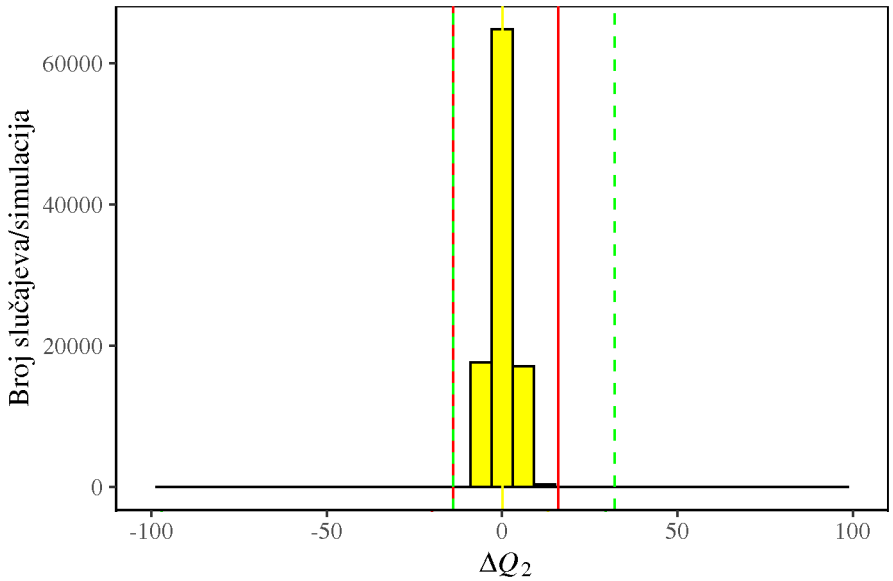
Iz podataka u Tablici 3.1 vidljivo je da je srednja vrijednost svih simulacijskih vrijednosti parametra Q_2 jednaka nasumičnoj točnosti $Q_{2,rnd}$ dobivenoj prema formuli (3.9) odnosno (3.10). Stoga, parametar $Q_{2,rnd}$ nazvan u [34] „najvjerojatnija nasumična točnost“, odgovara srednjoj vrijednosti svih mogućih vrijednosti parametra Q_2.

Stvarna točnost modela – ΔQ_2

Na *Slikama 3.3 i 3.4* prikazani su histogrami parametra ΔQ_2 iz kojih je razvidan utjecaj disbalansa klasa na neuravnoteženost klasa stvarne točnosti modela.



Slika 3.3 Raspodjela parametra kvalitete ΔQ_2 za varijable E i M s omjerom klasa 50:50 %



Slika 3.4 Raspodjela parametra kvalitete ΔQ_2 za varijable E i M s omjerom klasa 80:20 %

Na *Slikama 3.3 i 3.4* razvidno je kako stvarna točnost ima veći raspon kod uravnoteženog skupa (50:50 %). Na slikama je također razvidna promjena raspona promjenom udjela klase 1 (x), odnosno povećanjem neuravnoteženosti klasa. U *Tablici 3.2* dani su iznosi karakterističnih vrijednosti za dva omjera klase 1 u varijablama E i M sa *Slika 3.3* i *3.4*.

Tablica 3.2 Usporedba simulacijskih vrijednosti parametra ΔQ_2 za različite omjere klase 1

	$\Delta Q_{2,50}$	$\Delta Q_{2,80}$
Minimum	-22	-8
Maksimum	22	16
Medijan	0	0
Srednja vrijednost	0.016	0.007
Apsolutni minimum	-50	-8
Apsolutni maksimum	50	32

Kod parametra ΔQ_2 došlo je do promjene u apsolutnom i u simulacijskom rasponu koji je jednak 44 % odnosno 24 %, što znači da raspon stvarne točnosti ΔQ_2 opada porastom neuravnoteženosti klasa. Analogan zaključak vrijedi za apsolutne vrijednosti $\Delta(\Delta Q_2)$, tj. da se stvarni najveći mogući doprinos modela ΔQ_2 smanjuje povećanjem neuravnoteženosti udjela klasa.

3.1.3 Usporedba simulacijskih i izvedenih vrijednosti parametra točnosti

Kako bi se vrijednosti dobivene simulacijama usporedile s vrijednostima prema izvedenim formulama u širokom rasponu vrijednosti udjela klase 1 (x), provedene su detaljnije simulacije u rasponu udjela klase 1 od 0.01 do 0.99 (od 1 % do 99 %). Udjeli klase 1 jednaki 0 i 1 (0 % i 100 %) nije korišten u simulaciji zbog koji nije definiran u tim slučajevima i nije ga moguće jednostavno izračunati. Osim parametara $Q_2, \Delta Q_2$ i $Q_{2,rnd}$ koji će biti analizirani u ovom poglavlju, u projektu „simulator" skriptom „simetrijskeStandalone.R" (*Prilozi 3.41*) računaju se još neki dodatni parametara, i oni će biti spomenuti u nastavku.

Rezultati simulacija grafički su prikazani za minimalne i maksimalne vrijednosti, te za raspone. Izvedene karakteristične vrijednosti parametara zbirno su dane u *Tablici 3.1* u ovisnosti o udjelu klase 1. Maksimalna vrijednost parametra prikazana je jednom funkcijom u cijelom rasponu omjera klase 1 (od 0 do 1), a minimalna vrijednost dvjema funkcijama u pod-intervalu $x \in [0,1/2]$, a druga za $x \in [1/2,1]$. U slučajevima kada parametar nema rubne vrijednosti, segmenti su zamijenjeni intervalima.

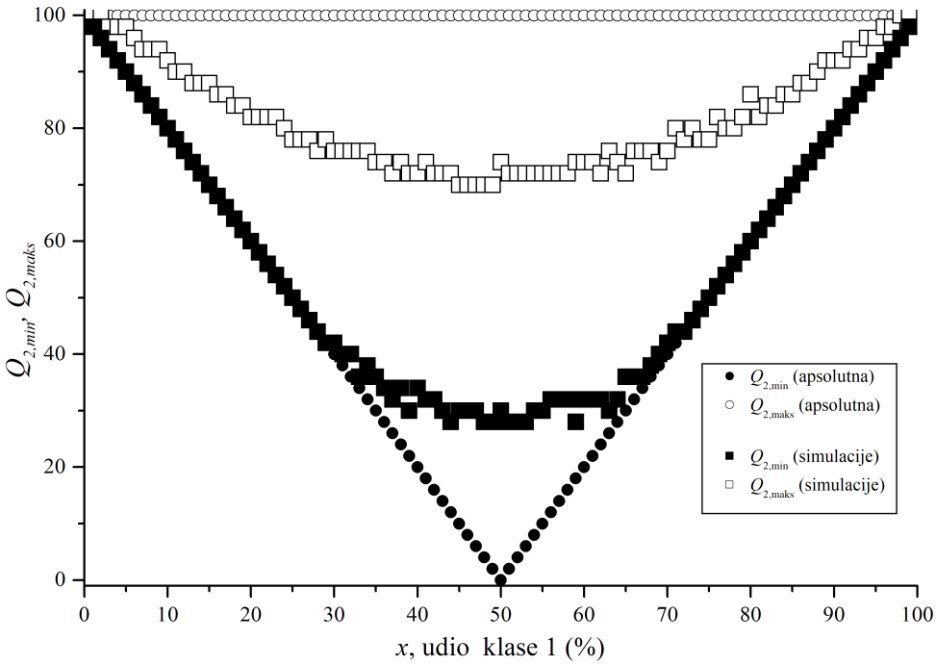
Koristeći skriptu "checkMinMaxFormulae.R" koja je dio simulatora (*Prilog 3.41*) korištena je za usporedbu karakterističnih vrijednosti parametara dobivenih simulacijom (apsolutne vrijednosti), te istih parametara dobivenih izvedenim formulama (teorijske/izvedene apsolutne vrijednosti). Izvedene formule različitih parametara u ovisnosti o udjelu klase 1, nalaze se u *Tablici 3.3*.

Tablica 3.3. Formule karakterističnih vrijednosti izmjenjivih varijabli za sve parametre ovisne o udjelu klase 1 i načinu preslagivanja podataka

	Formule karakterističnih vrijednosti parametara		
Poredak podataka	*AA*	*AD*	*AD*
Parametri	$x \in [0,1]$	$x \in [0,\frac{1}{2}]$	$x \in [\frac{1}{2},1]$
Q_2	1	$1-2x$	$2x-1$
$Q_{2,rnd}$	$2x^2-2x+1$	$2x^2-2x+1$	$2x^2-2x+1$
ΔQ_2	$-2x(x-1)$	$-2x^2$	$-2(x-1)^2$

***** *AA* – identično poredane vrijednosti varijabli E i M; *AD* – suprotno poredane vrijednosti varijabli E i M; x – udio klase 1

Svaki od parametara ima svoje simulacijske i apsolutne vrijednosti, pri čemu su simulacijske definirane kao one dobivene usporedbom podudarnosti vrijednosti varijabli E i M slučajnim izborom sadržaja varijable M bez njenog sortiranja. Suprotno tome, apsolutne minimalne i maksimalne simulacijske vrijednosti dobivene su tako da su vrijednosti varijabli E i M preslagane na jedan od dva načina, tj. ili obje varijable u identičnom poretku, ili obje varijable u obrnutom poretku. Minimalne i maksimalne vrijednosti parametra Q_2 u ovisnosti o udjelu klase 1 u rasponu od 1 do 99 %, prikazane su na grafu (*Slika 3.5*).



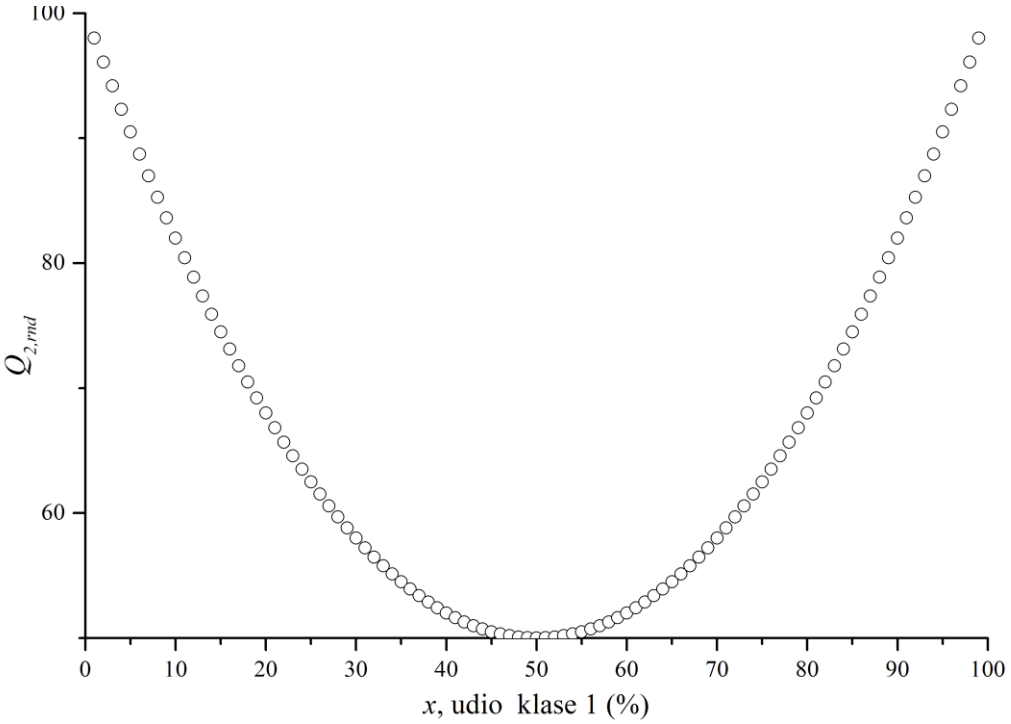
Slika 3.5 Apsolutne i simulacijske karakteristične vrijednosti parametra $Q_2(\%)$ u ovisnosti o udjelu klase 1 (x)

Na grafu (*Slika 3.5*) prikazane su apsolutne i simulacijske minimalne i maksimalne vrijednosti parametra Q_2. Uočava se kako minimalna simulacijska vrijednost nikada nije manja od minimalne apsolutne vrijednosti. Isto tako, maksimalna simulacijska vrijednost nikada nije veća od

maksimalne apsolutne vrijednosti. Ti rezultati sa *Slike 3.5* potvrđuju ispravnost algoritma za provedenu simulacija te, istovremeno, ispravnost izvedenih izraza za izračun karakterističnih vrijednosti parametara točnosti Q_2, odnosno stvarnog doprinosa modela iznad nasumične točnosti (ΔQ_2). Do poklapanja vrijednosti došlo je na krajevima gdje se približno 30% simulacijskih minimuma s obje strane izjednačilo s apsolutnim minimumima. U dijelu gdje nije došlo do poklapanja razlika je to veća što je udio klase 1 (x) bliži centru ($x = 50$ %). Razlog tome je što sa 100.000 simulacija za udio klase 1 jednak 50% nisu mogle biti realizirane sve moguće permutacije varijable M - jer je u tom slučaju (za $N = 100$) broj mogućih permutacija ogroman (~10^{29}). Stoga, za udjele klase 1 oko 50 % proanaliziran je samo jako mali dio (< 10^{-20} %) ukupnoga broja mogućih permutacija varijable M. Kako broj mogućih permutacija raste približavanjem središtu ($x = 50$ %), uz konstantan broj simulacija za taj udio klase 1 (uvijek jednak 100.000), jako je mala vjerojatnost za pronalazak permutacije varijable M koja bi dala minimalne i maksimalne karakteristične vrijednosti.

Ovdje je važno spomenuti da se maksimalna vrijednost parametra Q_2 uvijek može dobiti simulacijama, tj. sortiranjem varijabli E i M u identičnim poretcima, te ona će iznositi 1 (odnosno 100 %). Funkcionalna ovisnost minimalne apsolutne vrijednosti parametra Q_2 o udjelu klase 1 (x) je linearna na dva pod-intervala ($x \leq 50$ %, odnosno $x \geq 50$ %), što vidljivo na *Slici 3.5*. Uočava se da su jednadžbe tih pravaca $1 - 2x$ (za lijevi pod-interval udjela x) i $2x - 1$ (za desni pod-interval), i identične su izvedenim ovisnostima danim jednadžbama (3.1) i (3.2).

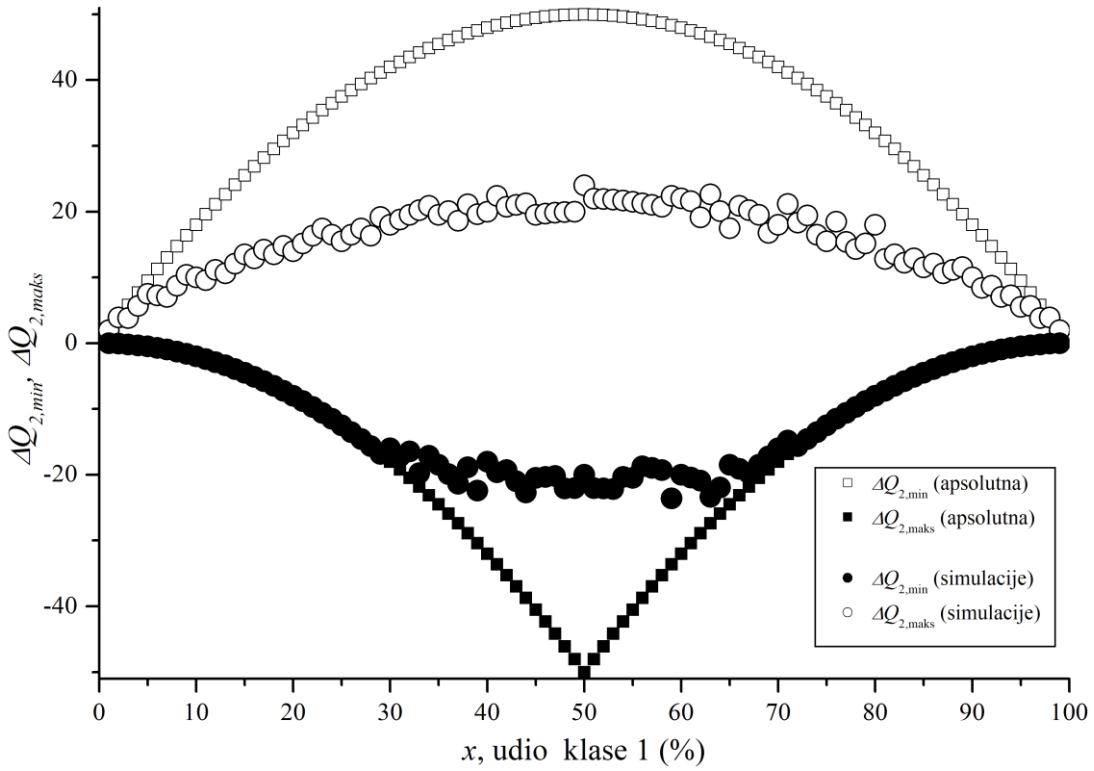
Ovisnost prosječne nasumične vrijednosti ($Q_{2,rnd}$) parametra Q_2 o udjelu klase 1 (x), prikazana je u grafu na *Slici 3.6.*



Slika 3.6 Vrijednosti parametra $Q_{2,rnd}$ u ovisnosti o udjelu klase 1 (x

Za parametar $Q_{2,rnd}$ pokazalo se da njegova vrijednost ne ovisi o poretku podataka unutar varijable nego samo o udjelu klase 1 u izmjenjivim varijablama. Funkcionalna veza između

udjela klase 1 i $Q_{2,rnd}$ (tj. i njegove maksimalne, minimalne i prosječne vrijednosti) je kvadratna, što potvrđuje ispravnost izvoda koji je dan jednadžbom (3.11) ($2x^2 - 2x + 1$). Odnos apsolutnih i simulacijskih vrijednosti parametra ΔQ_2 prikazan je na grafu (*Slika 3.9*).

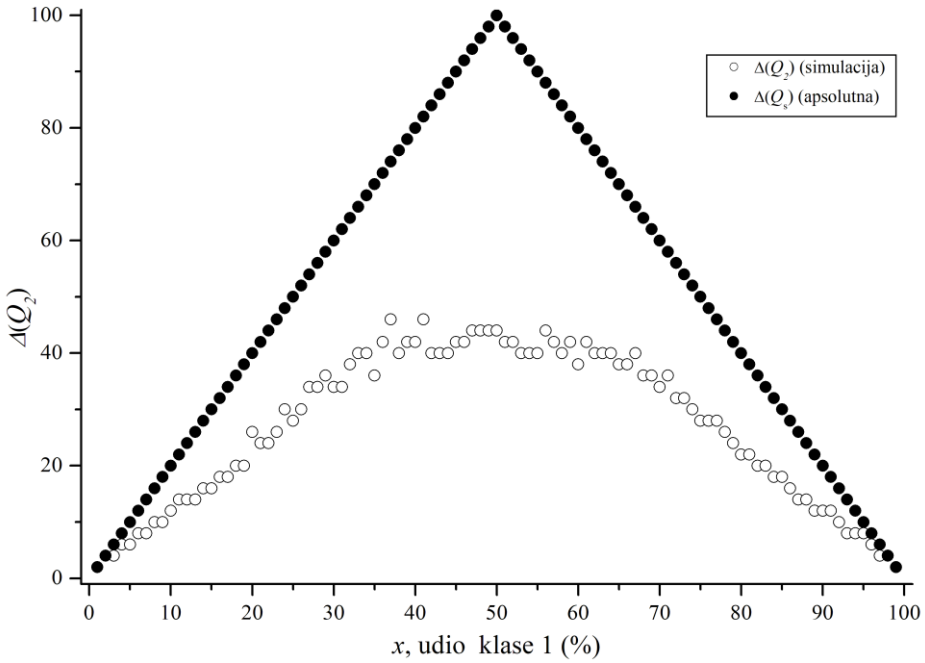


Slika 3.7 Apsolutne i simulacijske karakteristične vrijednosti parametra ΔQ_2 u ovisnosti o udjelu klase 1 (x)

Pokazalo se da simulacijske minimalne i maksimalne vrijednosti ΔQ_2 nikad ne prelaze iznad maksimalnih ili ispod minimalnih apsolutnih vrijednosti. Maksimalna vrijednost parametra ΔQ_2 u ovisnosti o udjelu klase 1 (x) prikazana je parabolom s negativnim predznakom prema formuli (3.15) i *Tablici 3.3*, jednako kao i za minimalne vrijednosti koje su također parabole s negativnim predznakom. Taj rezultat potvrđuje da su ispravni izrazi izvedeni za karakteristične vrijednosti parametra stvarne točnosti (točnosti koja je iznad nasumičnog pogađanja) ΔQ_2.

I kod ovog parametra, kao i kod Q_2 postoji dosta točaka gdje su apsolutne maksimalne vrijednosti jednake simulacijskim, i to je slučaj samo kod malih (< 30 %) ili velikih (> 70 %) udjela klase 1. Razlog za to nalazi se u činjenici da se kod manjih udjela x s 100.000 permutacija varijable M u simulacijama pronađu sve kombinacije, uključujući i one koje daju minimalnu i maksimalnu vrijednost parametra ΔQ_2.

Radi provjere ispravnosti izvedenih izraza za raspon između maksimalne i minimalne vrijednosti parametra Q_2 koja je označrna s $\Delta(Q_2)$, prikazane su na grafu (*Slika 3.8*) simulacijske i apsolutne vrijednosti raspona u ovisnosti o udjelu klase 1 (x) od 1 % do 99 %.



Slika 3.8 Apsolutni i simulacijski rasponi parametra $\Delta\ Q_2$ u ovisnosti o udjelu klase 1 (x)

Rasponi $\Delta\ Q_2$ i $\Delta\ \Delta Q_2$ jednaki su, što je i algebarskim putem dokazano formulama (3.7) i (3.8) ili malo drugačije napisano jednadžbom (3.18):

$$\Delta\ Q_2\ = \Delta(\Delta(Q_2)) = 100 - |100 - 2x|, \forall x \in [0{,}100] \tag{3.18}$$

Iz te jednakosti također zaključujemo da minimalna i maksimalna vrijednost parametra $\Delta\ Q_2$ nije ovisna o permutaciji varijable M. Osim toga, funkcijska ovisnost raspona o udjelu klase 1 ima oblik $2x$ za $x \leq 50$ %, i $2(1-x)$ za $x \geq 50$ % (jedn. (3.7)). Ta se ovisnost potvrđuje i na grafu (*Slika 3.8*) (puni crni kružići).

3.2. Entropija (složenost) varijable

Pojam entropije najviše se veže uz fiziku odnosno termodinamiku i statističku fiziku. Definira se za određeni fizikalni sustav - kao mjera neuređenosti (odnosno uređenosti) sustava s kojom je energija u njemu uskladištena. U fizikalnim sustavima entropija ovisi o njegovoj temperaturi, a promjene entropije povezane su s izmjenom topline između sustava i okoline. Pritom, u realnim sustavima izmjena topline ovisi o putu, a označava se velikim slovom S. Entropija je funkcija stanja fizikalnog sustava, i ne može se izračunati apsolutno, nego se samo mogu pratiti i računati njezine promjene između dva stanja sustava ΔS. Prema drugom zakonu termodinamike, za entropiju sustava termički izoliranih od okoline uvijek vrijedi: $\Delta S \geq 0$, pri čemu se znak jednakosti odnosi na povratne a znak nejednakosti za nepovratne procese.

Međutim, koncept entropije koristi se i u brojnim drugim područjima, poput teorije informacija [69,70]. Količina informacije u nekom sustavu mjeri se izračunavanjem entropije te informacije. Intuitivno je jasno da sustav koji sadrži veću količinu informacije ima i veću složenost. Zapravo, pojednostavljeno rečeno, količina informacije u sustavu izražava se preko njegove negativne entropije [69,70]. Ako promatramo varijablu s N vrijednosti koje mogu poprimati samo iznose 1 ili 0 (klasa 1 i klasa 0), jasno je da ukupni broj vrijednosti (elemenata) klase 1 (X) i klase 0 ($N - X$)

definira količinu informacije sadržane u toj varijabli. Međutim, pitanje je - kako ispravno kvantificirati tu količinu informacije? U tome pomaže koncept permutacije vrijednosti varijable, koji izračun količine informacije u varijabli svodi na prebrojavanje svih mogućih neidentičnih permutacija vrijednosti varijable.

Takav pristup ekvivalentan je konceptu Boltzmannove entropije u statističkoj fizici, koji ju računa na temelju particijske funkcije (W), koja prebrojava sva moguća energetski neidentična stanja sustava. Kako se u fizici u pravilu radi o velikom broju atoma ili molekula u sustavu, a koncept entropije najprije je razmatran na idealnim plinovima, kao jedini praktični način definiranja entropije. Iznos entropije tako je proporcionalan logaritmu particijske funkcije. Konstanta proporcionalnosti je Boltzmannova konstanta (k), pa je dobro poznati izraz za izračun entropije sustava $S = k \cdot lnW$.

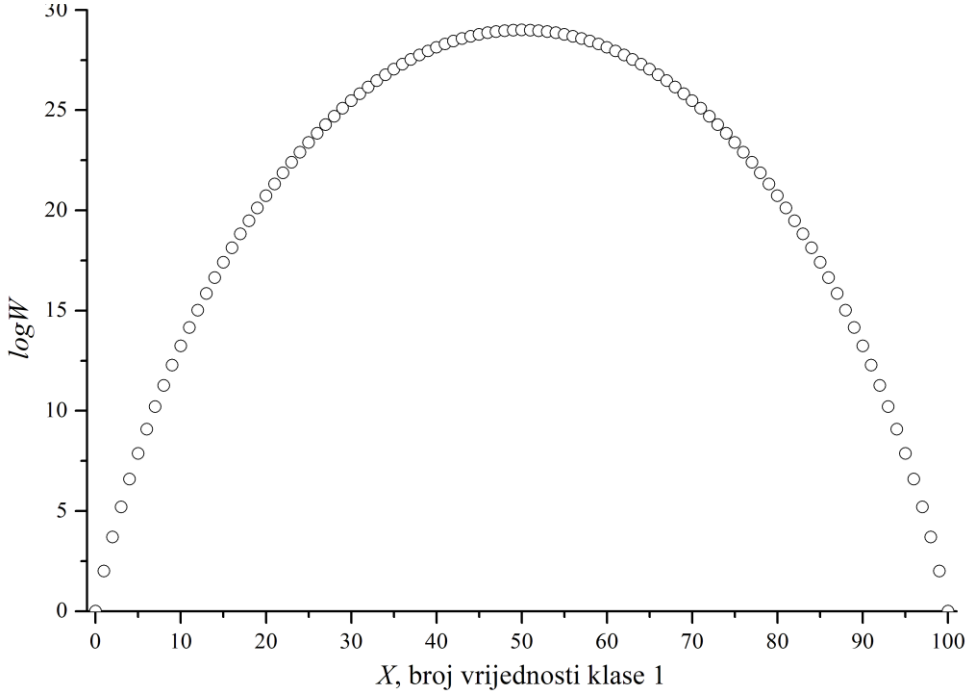
Particijska funkcija W, odnosno broj neidentičnih permutacija varijable koja ima N vrijednosti od kojih je njih X ima vrijednost 1 (klasa 1) a preostalih $(N - X)$ vrijednost 0 (klasa 0) računa se izrazom (3.19)

$$W = \frac{N!}{X!\,(N-X)!} = \binom{N}{X} \tag{3.19}$$

Zanemarujući fizikalnu konstantu k jer se ne radi o fizikalnom sustavu, entropija varijable računa se po Boltzmanovoj formuli (3.20):

$$S \sim log_{10} W = log_{10} \frac{N!}{X!\,(N-X)!} = log_{10} \binom{N}{X} \tag{3.20}$$

Pritom rabimo dekadski logaritam kao praktičniji jer se lakše može predočiti red veličine stvarnog broja permutacija varijable. Ovisnost dekadskog logaritma entropije varijable izračunane prema jednadžbi (3.20) i ukupnog broja vrijednosti klase 1 u varijabli (X) u rasponu od 0 do 100 dan je na grafu (*Slika 3.9*):



Slika 3.9. Ovisnost logaritma entropije $logW$ o broju vrijednosti klase 1 (X) u varijabli

Uočava se vrlo slična ovisnost kao ona koja je dobivena na grafu prikazanom na *Slici 3.6* između prosječne nasumične točnosti $Q_{2,rnd}$ i udjela vrijednosti klase 1 u varijabli (x). Entropija varijable s N vrijednosti raspoređene u samo dvije klase (1 i 0) ima to veću vrijednost što je vrijednost veličine X (broj elemenata jedne klase 1) u nazivniku formule (3.20) bliži $N/2$, a to se vidi i na grafu *Slike 3.9*. Vrijednost entropije varijable $logW$ koja ima tri ili više klasa (tri ili više identičnih vrijednosti) $X1$, $X2$, $X3$, ..., računao bi se prema jedn. (3.21):

$$logW = \log_{10} \frac{N!}{X1!\,X2!\,X3!\,\ldots} \tag{3.21}$$

Izračun faktorijela u formuli za entropiju zahtjevan je ukoliko je riječ o varijablama s velikim brojem vrijednosti N, i nije ih moguće izračunati i prikazati standardnim računalima. U radu s programom Excel, problemi u računanju faktorijela nastupaju već kod $N = 171$. Stoga se u računanju faktorijela velikog broja n koristi Stirlingova aproksimacija dana jednadžbom (3.22) [71]

$$log(n!) = nlog(n) + n + O(ln(n)) \tag{3.22}$$

gdje $O(ln(n))$ iznačava ostatak (korekciju) koji se za velike n može zanemariti. U većini analiza ovisnosti karakterističnih parametara i njihovih raspona s entropijom u disertaciji, rezultati su dani za 99 varijabli od kojih je svaka s ukupno $N = 100$ vrijednosti, a raspon ukupnog broja vrijednosti klase 1 (X) u varijablama kreće se od 1 do 99. Za te vrijednosti N i X entropija je računana prema jednadžbi (3.20), stoga što bi za male X izračun prema jedn. (3.22) unio veliku pogrešku. Rezultat Strilingove aproksimacije (3.22) je i formula za entropiju (3.23) s logaritmom po bazi 2:

$$H(x) = -(xlog_2(x) + (1-x)\log_2(1-x)) \tag{3.23}$$

koja se još naziva i binarna entropija, i označava se s $H(x)$. U jednadžbi (3.23) veličina $x = X/N$ predstavlja udio klase 1. Izraz (3.23) jako često koristi se u teoriji informacija, osobito danas kada se u pravilu analiziraju informacije u binarnom računalnom zapisu. U disertaciji će se koristiti dekadski logaritmi, pa je i izraz za informacijsku entropiju (3.23) prikazan u toj bazi (3.24):

$$H(x) = -(xlog_{10}(x) + (1-x)\log_{10}(1-x)) \tag{3.24}$$

Za veliki N izraz $logW$ rezultira vrijednostima koje je teško interpretirati, te je korisno imati normaliziranu vrijednost. Normalizacija se provodi vrijednošću entropije $logW$ najsloženije varijable za dani N, a to je varijabla koja ima $N/2$ vrijednosti koje su jednake 1 (a preostalih $N/2$ vrijednosti automatski je jednako 0). Vrijednost entropije najsloženije moguće varijable s N vrijednosti označena je s $logW_{Max}$, a normalizirana entropija računa se onda prema izrazu (3.25):

$$logW_{norm} = \frac{logW}{logW_{Max}} = \frac{\log_{10} \frac{N!}{X!\,(N-X)!}}{\log_{10} \frac{N!}{\frac{N}{2}!\,\frac{N}{2}!}} \tag{3.25}$$

Uz pomoć normalizirane vrijednosti entropije moguće je uspoređivati složenost varijabli različitog broja podataka koji pripadaju klasi 1 (X), Ukoliko se normalizirana entropija prema (3.25) za varijablu s ukupno N vrijednosti iskaže u postotcima, moguće je odrediti postotnu složenost svake varijable s N vrijednosti (za bilo koji omjer ukupnog broja klasa $X/(N-X)$) u odnosu na najsloženiju moguću varijablu. Također, moguće je za svaku varijablu izračunati pragove postotne složenosti iznosa 1 %, 5 %, 10 %, ili bilo koji drugi postotni prag, te ga koristiti za izbor varijabli

minimalne prihvatljive složenosti, Naime, u ranoj fazi modeliranja, unaprijed bi se odredio minimalni postotni prag složenosti za deskriptore od kojih bi se izabirali deskriptore u modele. Deskriptori niže složenosti bili bi isključeni iz postupka modeliranja.

Funkcije zaokruživanja na prvi veći i prvi manji cijeli broj korištene su u nazivniku formule (3.25) zbog varijabli s neparnim brojem elemenata. Ukoliko se te funkcije ne bi ispravno koristile, ne bi bile ispravno izračunane maksimalne vrijednosti entropije. Naime, algoritam za računanje logaritma faktorijela računao bi faktorijele decimalnih brojeva, dalo krivi rezultat. Dekadski logaritam u formulama odabran je radi lakše interpretacije rezultata, iako bi i odabir bilo koje druge baze dao jednako valjane ovisnosti i relacije između entropije i drugih parametara. Izvorni kod aplikacije Simulator razvijen u disertaciji dan je u *Prilogu 3.41*, a entropija je u njemu pohranjena u varijabli oznake „$entX$".

Algoritam za računanje složenosti implementiran u mrežnu aplikaciju (server) ima mogućnost računanja prilagođene Boltzmannove entropije (formula (3.21)) jer je taj izraz prilagođen i za varijable s jako malo vrijednosti (mali N) [72]. Naime, u slučajevima malih vrijednosti N, pogreška u primjeni Stirlingove aproksimacije prema (3.22) bila bi značajna. U mrežnom serveru implementirane su obje verzije entropije. S „complexity" označen je parametar $logW$, a normirana entropija označena je s „complexityNorm". Sve formule na mrežnom serveru vrijede za slučajeve kada u varijabli postoje samo dvije klase. Međutim, ukoliko varijabla ima više od dvije različite vrijednosti, uporabom dihotomizacije varijable moguće je dobiti aproksimativnu procjenu složenosti varijable.

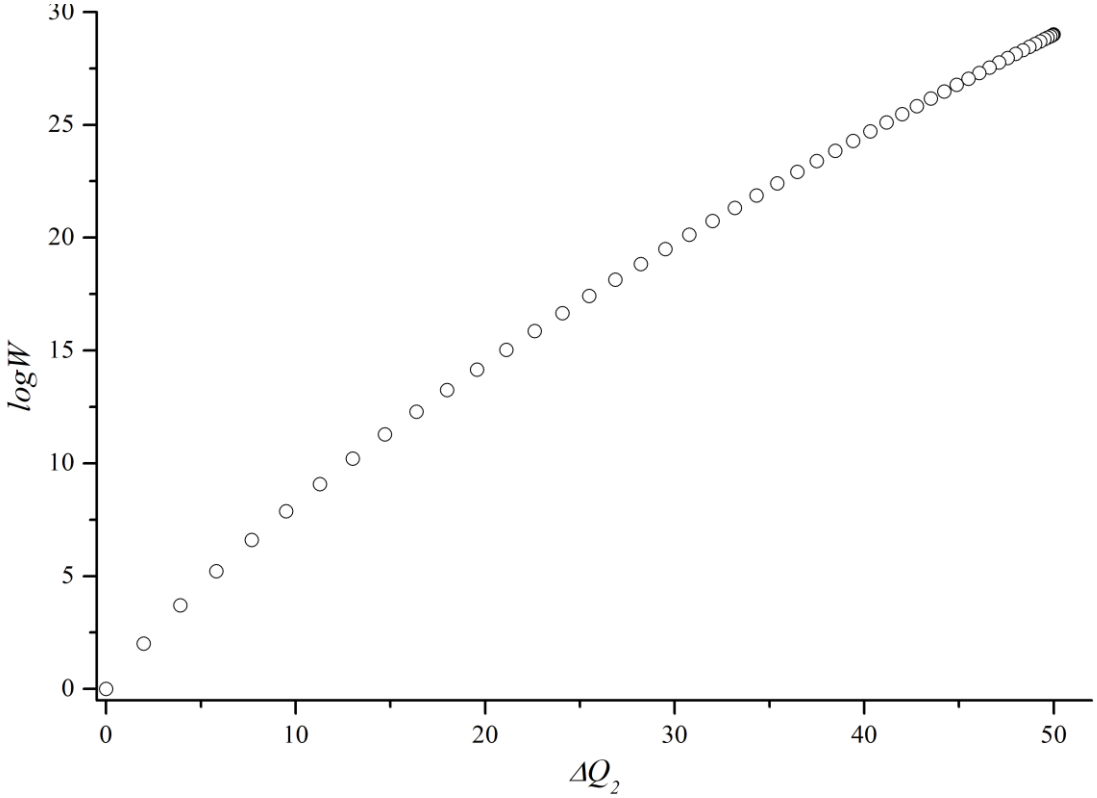
3.3 Usporedba izvedenih karakterističnih vrijednosti parametara točnosti s entropijom

U poglavlju 3.1 pokazano je da karakteristične vrijednosti parametra točnosti Q_2 izračunane između izmjenjivih varijabli E i M, poput minimalne i prosječne vrijednosti te razlika između minimalne, maksimalne i prosječne vrijednosti, pokazuju jasnu ovisnost o udjelu klase 1 (x) u varijablama (ili, što je ekvivalentno, o broju vrijednosti klase 1 u varijabli (X)). Slično je pokazano na grafovima (*Slika 3.9*) za entropiju varijable u ovisnosti o (X). Osobito je uočljiva visoka sličnost između ovisnosti na *Slikama 3.6* i *3.9*. Stoga, zanimljivo je istražiti korelacije između Q_2, $Q_{2,rnd}$ i ΔQ_2 s jedne, i entropije s druge strane. U *Tablici 3.4* dani su koeficijenti korelacije između apsolutnih i simulacijskih vrijednosti parametara Q_2, $Q_{2,rnd}$ i ΔQ_2.Vrijednosti koje se koriste u tablici su dobivene iz simulacija (vrijednosti X-a od 1 do 99), te su grupirane i njihove minimalne i maksimalne simulacijske i apsolutne vrijednosti su korelirane s $logW$.

Tablica 3.4 Koeficijenti korelacija između apsolutnih i simulacijskih karakterističnih vrijednosti parametara točnosti s entropijom ($logW$)

	minimalni R (koef. korelacije)		maksimalni R (koef. korelacije)	
parametar	apsolutni	simulacijski	apsolutni	simulacijski
$Q_{2,min}$	-0.949	-0.985		-0.987
$Q_{2,rnd}$	-0.997	-0.997	-0.997	-0.997
ΔQ_2	-0.844	-0.942	0.997	0.980

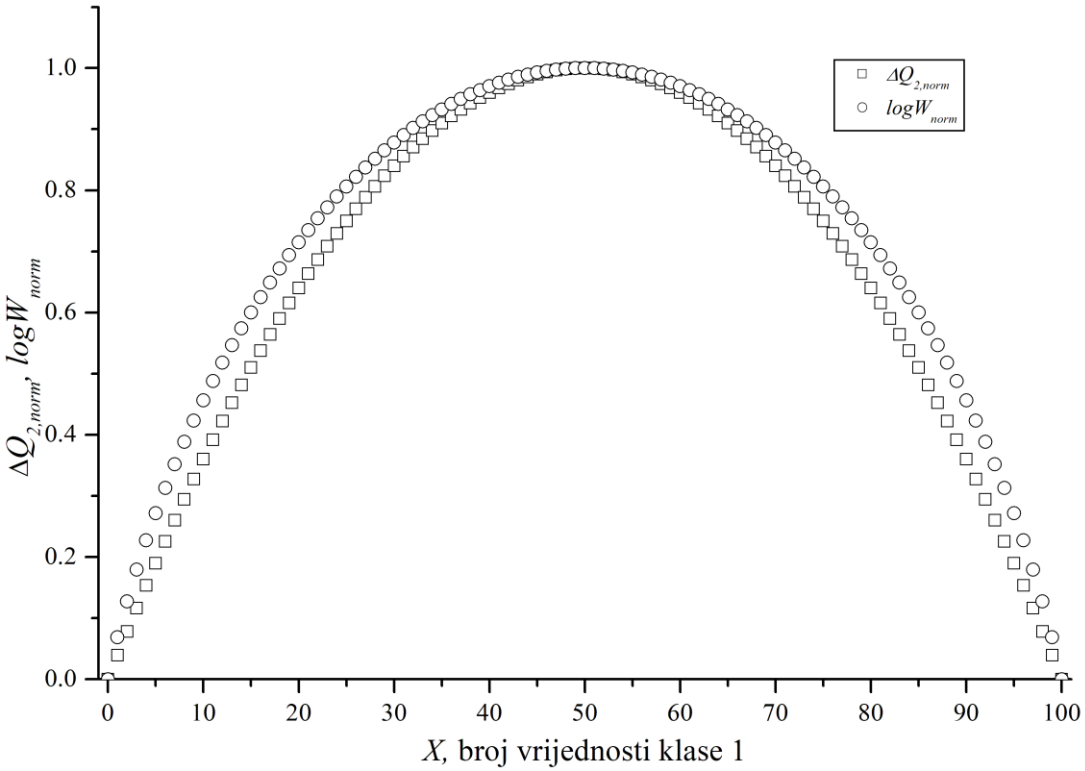
Najveći iznosi korelacije ($R = -0.997$) dobiveni su između entropije i prosječne nasumične točnosti $Q_{2,rnd}$. Identična je korelacija, samo pozitivnog predznaka, dobivena između entropije i apsolutnog doprinosa modela iznad prosječne nasumične točnosti ($\Delta Q_2 = 1 - Q_{2,rnd}$). Naime, za izmjenjive varijable, maksimalni Q_2 uvijek je jednak 1 (odnosno 100 %). Sve korelacije računane su na 99 vrijednosti u rasponu udjela klase 1 od 1 % do 99 %, analizirajući podudarnost (točnost, poklapanje) između izmjenjivih varijabli E i M, kako je opisano u poglavlju 3.2. Grafički prikaz ovisnosti (korelacije) između logaritma entropije ($logW$) i $\Delta Q_2 = 1 - Q_{2,rnd}$ za 101 varijablu u rasponu broja vrijednosti klase 1 (X) od 0 do 100 dan je na *Slici 3.10*:



Slika 3.10. Ovisnost između logaritma ($logW$) entropije i doprinosa modela iznad nasumične točnosti (ΔQ_2)

Prisutna je linearna ovisnost između $logW$ i (ΔQ_2), osim pojedinačnih odstupanja kada je broj vrijednosti klase 1 u varijabli (X) manji od (približno) 10 %.

Ovisnost između logaritma ($logW$) entropije i doprinosa modela iznad nasumične točnosti (ΔQ_2) izdvajaju se kao parametri koji najbolje koreliraju, te su stoga glavni kandidati za analizu složenosti varijable. Kako bi se procijenila složenost varijabli za dani N s različitim X, potrebno je normirati vrijednosti parametara vrijednošću najsloženije moguće varijable za dani N, a to je varijabla koja ima $X = N/2$ (ili, ukoliko je N neparan, X koji je najbliži vrijednosti $N/2$). Na *Slici 3.11* dana je ovisnost normirane vrijednosti entropije ($logW_{norm}$) i normirane stvarne točnosti iznad nasumične $\Delta Q_{2_{norm}}$.



Slika 3.11. Ovisnost normirane entropije ($logW_{norm}$) varijable i normiranog doprinosa modela iznad nasumične točnosti ($\Delta Q_{2_{norm}}$) o udjelu klasa u rasponu od 0 % do 100 %.

Parametar $\Delta Q_{2_{norm}}$ dobiven je analogno normaliziranoj entropiji, i predstavlja normalizirani stvarni doprinos modela koji je iznad nasumičnog pogađanja. Normalizacija se provodi varijablom koja ima maksimalni ΔQ_2, a to se događa kad varijabla s ΔQ_2 vrijednosti ima u svakoj klasi podjednak broj vrijednosti ($N/2$), kada je $\Delta Q_2 = 0.5$, odnosno 50 %. Analizom ovisnosti na *Slici 3.11* vidimo da je podudarnost visoka, s tim da je složenost iskazana preko $\Delta Q_{2_{norm}}$ uvijek nešto niža nego složenost iskazana normaliziranom entropijom. Stoga, ispada da je parametar ΔQ_2 nešto strožiji ako se koristi kao mjera za procjenu/iskazivanje složenosti klasifikacijske varijable.

3.4 Karakteristične vrijednosti dodatnih parametara *MAE*, *s*, *MCC*, *F1* i κ

Postupke simulacija i izvođenja matematičkih izraza primijenjene u poglavlju 3.1 za određivanje karakterističnih vrijednosti parametara točnosti Q_2 (tj. minimalnu, maksimalnu i prosječnu vrijednost točnosti, te njihove raspone) u ovom poglavlju koristit ćemo za određivanje karakterističnih vrijednosti drugih parametara koji se često rabe u analizama točnosti/pogreške modela. To su parametri koji mjere srednje apsolutnu (MAE) i standardnu pogrešku (s), te parametri koji mjere podudarnost/točnost između eksperimentalne (E) i modelne (M) varijable: koeficijent korelacije (MCC), *F1*-score ($F1$) i Cohenova kapa (κ). Kao i u ranijim simulacijskim analizama, i u ovom dijelu eksperimentalna varijabla E ima stalni (fiksni) poredak vrijednosti, dok je modelna varijabla M označava varijabla E permutirana u raznim porecima u odnosu na početni stalni poredak vrijednosti varijable E. Obje varijable imaju iste udjele klase 1 (x) i klase 0 ($1 - x$), i

par takvih varijabli predstavlja par izmjenjivih varijabli. U svakom od tih parova varijabli računa se pogreška (odstupanje) u njihovim vrijednostima ili njihova podudarnost, i iskazuje spomenutim parametrima.

Parametri *MAE*, *s*, *MCC*, *F1 i* κ imaju svoje simulacijske (približne) i izvedene (apsolutne) minimalne i maksimalne te prosječnu vrijednosti koje istražujemo u ovisnosti u udjelu klase 1 (x) u varijablama E i M. Možemo zamisliti da se permutiranjem varijable M u odnosu na varijablu E može dobiti i lošije i bolje poklapanje vrijednosti tih dviju varijabli, i da se kvaliteta poklapanja/podudarnosti varijabli može kvantificirati različitim vrijednostima parametara *MAE*, *s*, *MCC*, *F1 i* κ, što vodi na raspodjelu vrijednosti svakog od parametara. Raspodjele tih parametara i njihove karakteristične vrijednosti istražit će se u nastavku (u 3.4.1 i 3.4.2) u ovisnosti o udjelu klase 1 (x).

3.4.1 Izvodi karakterističnih vrijednosti parametara *MAE, s, MCC, F1* i κ

Minimalne i maksimalne karakterističnih vrijednosti parametara dobivene su koristeći tablicu supstitucija vrijednosti p, n, u i o iskazanih u ovisnosti o udjelu klase 1 (x) u varijabli i ukupnom broju vrijednosti u varijabli (N) (*Tablica 2.3*). Prosječne nasumične karakteristične vrijednosti dobivene su uvrštavanjem supstitucijskih izraza za srednje vrijednosti elemenata matrice pogrešaka p, n, u i o temeljene na teoriji vjerojatnosti [68] i na analogiji s izračunom parametra nasumične točnosti $Q_{2,rnd}$ (jednadžba (3.9) i literatura [34,35]).

Karakteristične vrijednosti prosječne apsolutne pogreške (*MAE*)

Općenito, prosječna apsolutna pogreška MAE računa se tako da se sve apsolutne razlike (u i o) između eksperimentalne E i modelne M varijable zbroje i podjele sa ukupnim brojem podataka u varijabli (N) prema jedn. (3.26)

$$MAE = \frac{o + u}{N} \tag{3.26}$$

a njezina prilagodba za izmjenjive varijable kod kojih je $u = o$ dana je izrazom (3.27):

$$MAE = \frac{2u}{N} \tag{3.27}$$

Detaljniji izvodi dani su u *Prilozima 3.13-3.17*.

Uvrštavajući supstitucijske relacije za u iz *Tablice 2.3* u izraz (3.27). dobivaju se relacije (3.28) do (3.34). Formula (3.28) odnosi se na određivanje maksimalne vrijednosti parametra MAE s pomoću obrnutog AD preslagivanja (sortiranja) varijabli E i M za lijevi (L) pod-interval ($x \leq 1\ \ 2$) u ovisnosti o udjelu klase 1 (x):

$$MAE_{AD,L} = 2x \tag{3.28}$$

Maksimalna vrijednost parametra MAE određuje se također obrnutim preslagivanjem AD za desni (R) pod-interval $x \geq 1\ \ 2$ (formula (3.29))

$$MAE_{AD,R} = 2(1 - x) \tag{3.29}$$

Minimalna vrijednost parametra MAE konstantna je u cijelom intervalu x, i njezina vrijednost je dana formulom (3.30)

$$MAE_{AA} = 0 \tag{3.30}$$

a do toga se dolazi jednakim AA preslagivanjem (sortiranjem) varijabli E i M. Korisno je pripomenuti da kod parametara koji mjere pogrešku, značenje minimuma i maksimuma obrnuta je u odnosu na parametar točnosti (podudarnosti) Q_2. Najveći mogući raspon MAE dan je formulama (3.31) do (3.33), a detaljniji izvodi dani su u *Prilozima 3.16 i 3.17.*

Maksimalna vrijednost MAE simetrična je u odnosu na vrijednost $x = 1\ \ 2$ što znači da za raspon MAE postoje dvije formule: (3.31) za $x \leq 1\ \ 2$ i (3.32) za $x \geq 1\ \ 2$.

$$\Delta(MAE_L) = 2x, \forall x \in [0, \frac{1}{2}] \tag{3.31}$$

$$\Delta(MAE_R) = 2(1-x), \forall x \in [\frac{1}{2}, 1] \tag{3.32}$$

Raspon parametra MAE u ovisnosti o udjelu klase 1 za cijelo područje x dan je formulom (3.33)

$$\Delta(MAE) = 1 - |1 - 2x|, \forall x \in [0,1] \tag{3.33}$$

Prosječna nasumična vrijednost MAE u ovisnosti o elementima matrice pogrešaka p, n, u i o prikazana je formulom (3.34).

$$MAE_{rnd} = \frac{2\ p+u\ \ n+u}{N^2} = \begin{matrix} o = u \\ x = \ p+u\ /N \\ 1-x = (n+u)/N \end{matrix} = 2x(1-x) \tag{3.34}$$

Karakteristične vrijednosti standardne pogreške *s*

Standardna pogreška s za klasifikacijske varijable s dva stanja ima vrlo sličnu formulu onoj kod srednje apsolutne pogreške MAE, s tom razlikom da je cijeli izraz pod korijenom. Izvodi izraza za karakteristične vrijednosti potpuno je analogan onom prethodno opisanom kod parametra MAE, a dani su izrazima (3.35) do (3.40):

$$s = \ \ 2\overline{\frac{u}{N}} \tag{3.35}$$

$$s_{AA} = 0, \forall x \in [0,1] \tag{3.36}$$

$$s_{AD,L} = \ \ \overline{2x}, \forall x \in [0, \frac{1}{2}] \tag{3.37}$$

$$s_{AD,R} = \ \ \overline{2(1-x)}, \forall x \in [\frac{1}{2}, 1] \tag{3.38}$$

$$\Delta(s_L) = \ \ \overline{2x}, \forall x \in [0, \frac{1}{2}] \tag{3.39}$$

$$\Delta(s_R) = \ \ \overline{2(1-x)}, \forall x \in [\frac{1}{2}, 1] \tag{3.40}$$

$$\Delta(s) = \ \ \overline{1-|1-2x|}, \forall x \in [0,1] \tag{3.41}$$

Izvodi formula (3.35) do (3.41) nalaze se u *Prilozima 3.18-3.23.*

Prosječna nasumična vrijednost s u ovisnosti o p, n, o, u parametrima dana je formulom (3.42).

$$s_{rnd} = \frac{\overline{2(p+o)(n+u)}}{N} = \begin{matrix} o = u \\ x = p+u\ /N \\ 1-x = (n+u)/N \end{matrix} = \overline{2x(1-x)} \tag{3.42}$$

Karakteristične vrijednosti koeficijenta korelacije *MCC*

Koristeći iste supstitucije za veličine p, n, u i o kao i za prethodne izvode (*Materijali i metode*, *Tablica 2.3*), dobivene su karakteristične vrijednosti parametra MCC u ovisnosti o *x*-u (udjelu klase 1). U formuli (3.43) dano je pojednostavljenje za izmjenjive varijable zbog $o = u$ (izvod se nalazi u prilozima – *Prilog 3.24*)

$$MCC = \frac{np - u^2}{(n+u)(p+u)} \tag{3.43}$$

Maksimalna vrijednost parametra MCC u ovisnosti o udjelu klase 1 prikazana je formulom **(**3.44) i ona je uvijek konstanta za izmjenjive varijable (izvod u *Prilogu 3.25*).

$$MCC_{AA} = MCC_{max} = 1, \forall x \in [0,1] \tag{3.44}$$

Taj je rezultat dobiven jednakim (AA) preslagivanjem vrijednosti varijabli E i M. Minimalne vrijednosti opisane su u ovisnosti o udjelu x izrazima (3.45) i (3.46).

$$MCC_{AD,L} = MCC_{min} = \frac{x}{x-1}, \forall x \in [0, \frac{1}{2}] \tag{3.45}$$

$$MCC_{AD,R} = MCC_{min} = \frac{x-1}{x}, \forall x \in [\frac{1}{2}, 1] \tag{3.46}$$

Izvod je dan u *Prilozima 3.26 i 3.27*. Razlika minimalnih i maksimalnih vrijednosti funkcija parametra MCC u ovisnosti o udjelu klase 1 predstavlja raspon parametra $\Delta\ MCC$ koji je prikazan jednadžbama (3.47) i (3.48).

$$\Delta(MCC_L) = \frac{1}{1-x}, \forall x \in [0, \frac{1}{2}] \tag{3.47}$$

$$\Delta(MCC_R) = \frac{1}{x}, \forall x \in [\frac{1}{2}, 1] \tag{3.48}$$

Prosječna nasumična vrijednost MCC uvijek je jednaka nuli:

$$MCC_{rnd} = 0 \tag{3.49}$$

a dobivena je uvrštavanjem supstitucijskih vrijednosti za veličine p, n i u (*Tablica 3.19*) u izraz za izračun parametra MCC u jednadžbu (2.14) i u *Prilogu 2.2*.

Karakteristične vrijednosti parametra *F1*

Ovaj parametar primarno se koristi za kvantificiranje kvalitete modela razvijenih na visoko neuravnoteženim skupovima podataka kod kojih je n jako velik, tj. puno veći od p, u i o [35].
Općeniti izraz za izračun $F1$ dan je jednadžbom (2.8), a uz pojednostavljenje zbog toga što je $u = o$ za izmjenjive varijable dobije se jedn. (3.50) (*Prilog 3.30*).

$$F1 = \frac{p}{p+u} \tag{3.50}$$

Za potrebe računanja minimalnih i maksimalnih vrijednosti parametra $F1$ u ovisnosti o udjelu klase 1, vrijednosti elemenata matrice pogrešaka p, u i o supstituirane su izrazima iz *Tablice 2.3 (Materijali i metode).* Svi se izvodi provode analogno kao i u slučaju parametara točnosti ili koeficijenta korelacije MCC. Maksimalna vrijednost parametra $F1$ konstantna je i jednaka 1 (3.51) a detalji su dani u *Prilogu 3.31*.

$$F1_{AA} = 1, \forall x \in \langle 0{,}1] \tag{3.51}$$

Minimalna vrijednost parametra $F1$ u lijevom pod-intervalu $x \in 0,1\ 2$ jednaka je nuli (formula (3.52) i *Prilog 3.32*):

$$F1_{AD,L} = 0, \forall x \in \langle 0, \frac{1}{2}] \tag{3.52}$$

Minimalna vrijednost parametra $F1$ u desnom pod-intervalu udjela klase 1 ($x \geq 1\ 2$), označena je s $F1_{AD,R}$, i njegova je vrijednost dana formulom (3.53):

$$F1_{AD,R} = \frac{2x-1}{x}, \forall x \in [\frac{1}{2}, 1] \tag{3.53}$$

Izvod je dostupan u *Prilogu 3.32*. Iz formula (3.52) i (3.53) vidljivo je da je riječ o asimetriji, tj. o parametru kojem lijeva i desna strana nisu jednake u odnosu na točku $x = 1/2$, što nije poželjno. Naime, zamjena oznaka klase prilikom unosa podataka može rezultirati potpuno drugačijim rezultatom i tumačenjem. Osim toga ovaj parametar nije definiran za slučaj $x = 0$.

Kada se oduzmu funkcije minimalnih i maksimalnih vrijednosti u ovisnosti o udjelu klase 1, dobije se raspon prikazan formulama (3.54) i (3.55).

$$\Delta(F1_L) = 1, \forall x \in \langle 0, \frac{1}{2}] \tag{3.54}$$

$$\Delta(F1_R) = \frac{1-x}{x}, \forall x \in [\frac{1}{2}, 1] \tag{3.55}$$

Iz formula (3.54) i (3.55) vidljivo je kako je raspon nesimetričan u odnosu na točku $x = 1/2$, tj. u rasponu $x \in \langle 0{,}1/2]$ vrijednost je 1, dok je u rasponu $x \in [1/2, 1]$ njegova vrijednost $(1-x)/x$. Izvod se nalazi u *Prilogu 3.34*.

Prosječna nasumična vrijednost $F1$ u ovistnosti o elementima matrice pogrešaka p, n, u i o prikazana je formulom (3.56). Izvod je dobiven uvrštavanjem izraza za srednje vrijednosti p, n, u i o temeljene na vjerojatnostima [68].

$$F1_{rnd} = \frac{p+u}{p+u\ +\ n+u} = \begin{matrix} X = p+u \\ N-X = (n+u) \\ x = X/N \end{matrix} = \frac{X}{X+(N-X)} = x \tag{3.56}$$

Karakteristične vrijednosti parametra Cohenove kape (κ)

Parametar Cohenov kapa (κ) korišten je za izračun stupnja slaganja među izmjenjivim varijablama E i M [38]. Izraz za njegov izračun za izmjenjive varijable identičan je onom za koeficijent korelacije MCC, što je pokazano jednadžbama (2.14) odnosno (3.43).

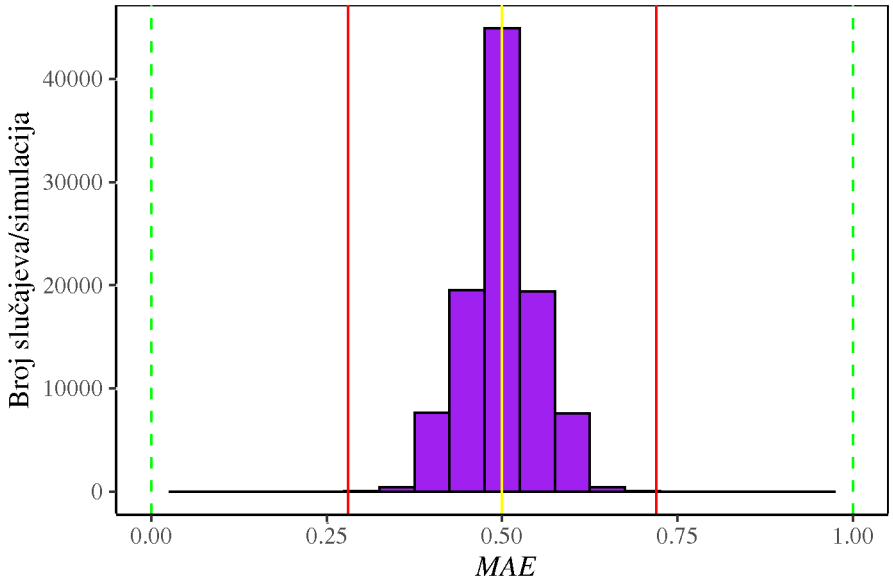
Stoga, izvodi karakterističnih vrijednosti i raspona ovog parametra posve su identični onima dobiveni za parametar MCC. Međutim, u općenitom slučaju kad je $u \neq o$, tj. kad nije riječ o usporedbi podudarnosti izmjenjivih varijabli i kad nije riječ o izračunu Cohenove kape za uravnoteženi model, karakteristične vrijednosti κ neće biti identične onima koeficijenta korelacije MCC.

3.4.2 Simulacije karakterističnih vrijednosti parametara MAE, s, MCC i $F1$

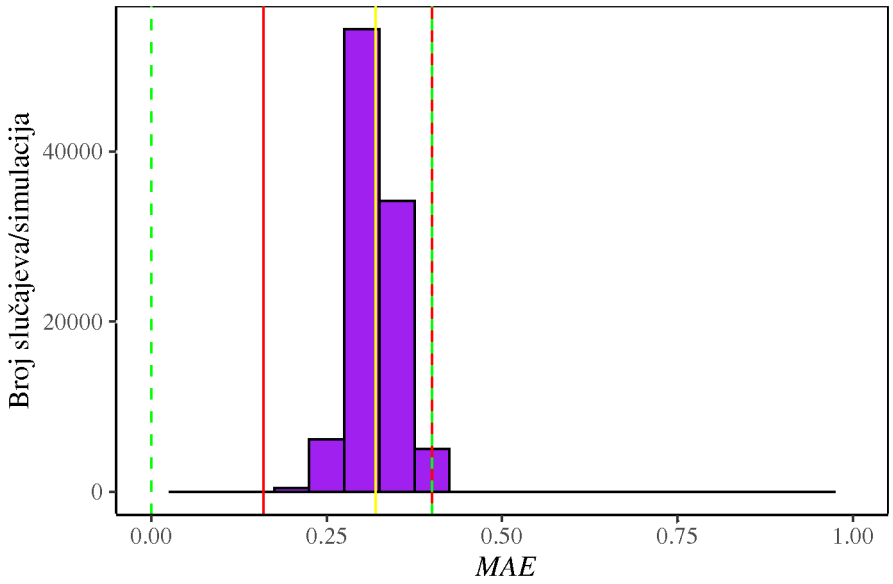
U svrhu potvrđivanja ispravnosti izvedenih izraza za određivanje minimalnih, maksimalnih i prosječnih vrijednosti parametara, te njihovih raspona, provedena su simulacijska istraživanja. Bit će prikazani rezultati za parametre MAE, s, MCC i $F1$ za dva udjela klasa (1 i 0), od kojih je jedan simetričan (50:50 %), dok je drugi udio klasa asimetričan (80:20 %). Svrha analize je i utvrditi utjecaj promjene udjela klase 1 na raspone parametara, te potom istražiti korelaciju raspona karakterističnih vrijednosti ovih parametara s entropijom varijable.

Raspodjela prosječne apsolutne pogreške – *MAE*

Raspodjela prosječne apsolutne pogreške za dva udjela klasa prikazana je grafovima (*Slike 3.12* i *3.13*). Crvene vertikalne crte predstavljaju simulacijske minimalne i maksimalne vrijednosti, dok zelene vertikalne crte predstavljaju apsolutne vrijednosti parametara dobivene odgovarajućim preslagivanjem varijabli. Kao i kod parametra točnosti, svaka simulacija provedena je na 100.000 parova izmjenjivih varijabli E i M.



Slika 3.12 Raspodjela prosječne apsolutne pogreške (*MAE*) varijable s omjerom klasa 50:50 %



Slika 3.13 Raspodjela prosječne apsolutne pogreške (MAE) dobivena simulacijama za varijable s omjerom klasa 80:20 %

Na *Slikama 3.12 i 3.13* vidljiv je pomak centra raspodjela kao i promjena širine raspodjela, što upućuje na to da promjena udjela klase 1 (x) utječe na raspon parametra MAE. Zbirni numerički podaci za ove raspodjele dani su u *Tablici 3.5*. Broj 50 u indeksu označava udio klase 1 $x = 50$ %, a broj 80 u indeksu označava udio klase 1 $x = 80$ %.

Tablica 3.5 Deskriptivni podaci raspodjela MAE za varijable s 50 % i 80 % udjela klase 1 ($x = 50$ % i $x = 80$ %)

	MAE_{50}	MAE_{80}
Minimum	0.2800	0.1600
Maksimum	0.7200	0.4000
Medijan	0.5000	0.3200
Srednja vrijednost	0.4998	0.3199
Apsolutni minimum	0	0
Apsolutni maksimum	1	0.4

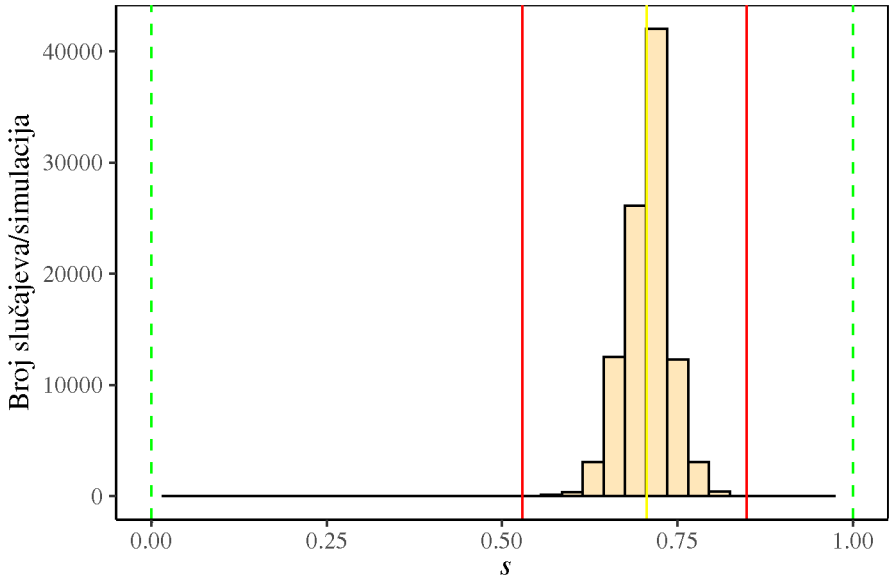
Kod prosječne apsolutne pogreške za udio 80:20 % došlo je do pomaka srednje vrijednosti raspodjele u lijevo u odnosu na raspodjelu za udjele klasa 50:50 % (0.32 u usporedbi s 0.59). Sličan pomak događa se i kod medijana.

Osim toga, porastom brojnosti klase 1 u odnosu na klasu 0, dolazi i do promjene apsolutnog raspona koji se smanjuje s 1 (skup 50:50 %) na 0.4 (skup 80:20 %). Analogno tome, simulacijski raspon također se smanjio s $\Delta(MAE_{50}) = 0.44$ na $\Delta(MAE_{80}) = 0.24$. Simulacijska i apsolutna vrijednost maksimuma preklopile su se za skup s udjelom klasa 80:20 % (vrijednost 0.4).

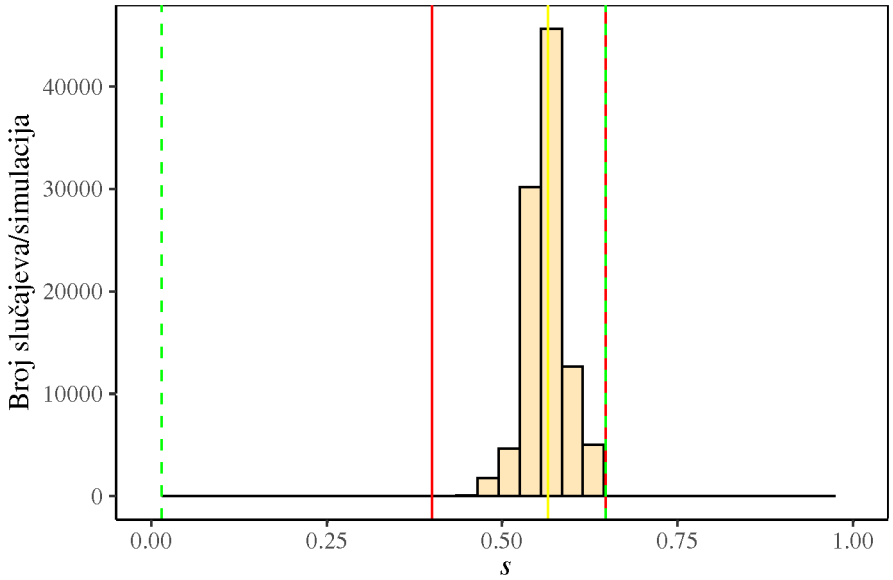
Raspodjela standardne pogreške – *s*

U svrhu promatranja utjecaja promjene udjela klase 1 na ponašanje standardne pogreške s, simulacijom dobivene vrijednosti prikazane su grafički na *Slikama 3.14 i 3.15*. Crtom žute boje

označena je srednja vrijednost raspodjela, crvenim vertikalnim crtama minimalne i maksimalne simulacijske vrijednosti, dok su zelenim vertikalnim crtama označene apsolutne vrijednosti.



Slika 3.14 Raspodjela standardne pogreške (s) za varijable s omjerom klasa 50:50 %



Slika 3.15 Raspodjela standardne pogreške (s) za varijable s omjerom klasa 80:20 %

Iz *Slika 3.14 i 3.15* vidljivo je da se kod nesimetričnog udjela klasa u varijabli (kod neuravnoteženosti između klasa) smanjuje raspon parametra s. Vidljivo je da su za slučaj $x = 80$ % smanjene maksimalna apsolutna i simulacijska vrijednost parametra s, što je zbirno prikazano u *Tablici 3.6.*

Tablica 3.6 Deskriptivni podaci raspodjela parametra *s* za simulacijske varijable s udjelima klasa 50:50 % i 80:20 %

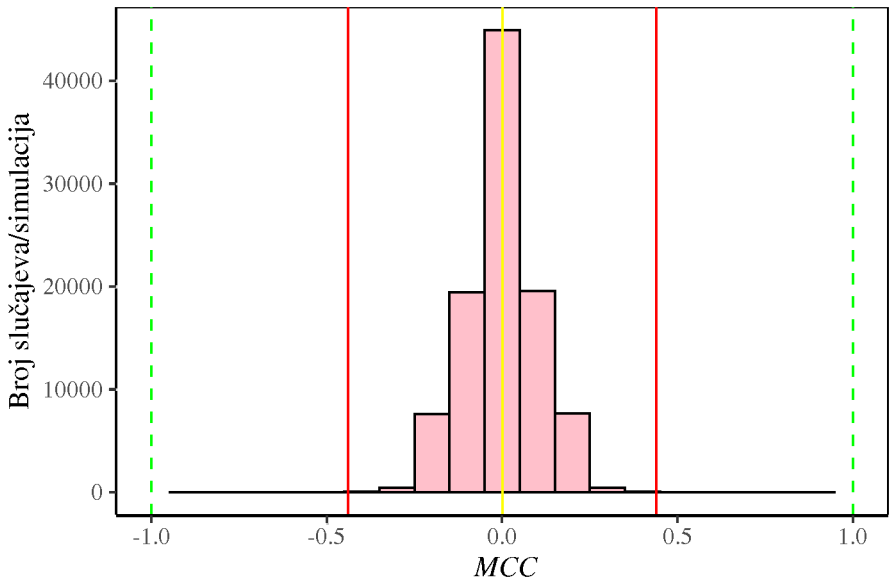
	s_{50}	s_{80}
Minimum	0.5292	0.4000
Maksimum	0.8485	0.6325
Medijan	0.7071	0.5657
Srednja vrijednost	0.7061	0.5649
Apsolutni minimum	0	0
Apsolutni maksimum	1	0.6325

Kod standardne pogreške raspon apsolutnih vrijednosti je od 0 do 1 kod simetričnog skupa 50:50 %, dok je kod nesimetričnog skupa 80:20 % raspon smanjen i on je od 0 do 0.6325. Simulacijski maksimum za raspodjelu 80:20 % je 0.4 što je za 0.2325 manje od apsolutnog maksimuma. Nadalje, simulacijski raspona iznosi 0.31. Srednja vrijednost kod simulacije sa simetričnim udjelima klasa iznosi 0.7061, dok je u skupu 80:20 % pomaknuta ulijevo i iznosi 0.5649.

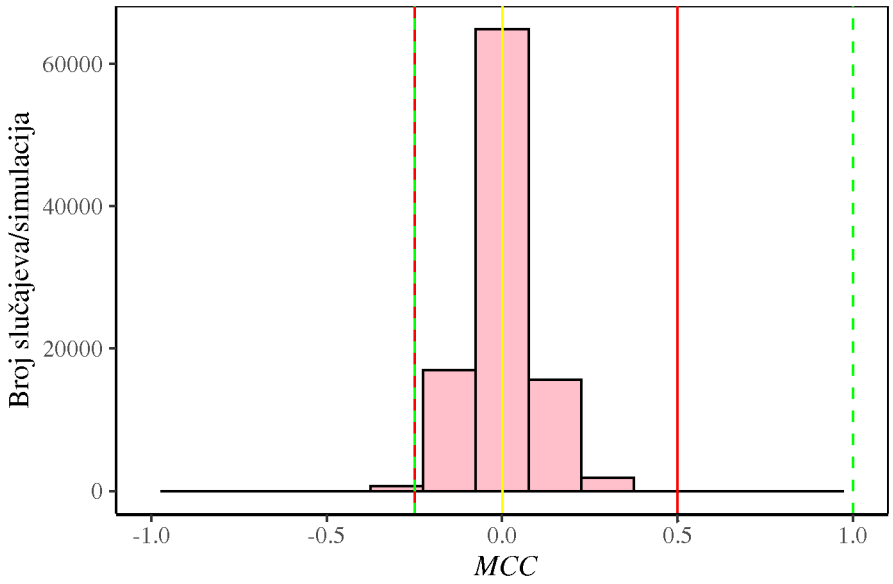
Apsolutni raspon raspodjele standardne pogreške za varijablu sa simetričnim udjelima klasa iznosi 1, dok za raspodjelu kod varijable s nesimetričnim udjelima raspon vrijednosti iznosi 0.63. Te promjene u raspodjeli upućuju na funkcionalnu ovisnost raspona standardne pogreške o udjelu jedne klase u varijablama.

Raspodjela Matthews-ovog koeficijenta korelacije – *MCC*

Matthews-ov koeficijent korelacije – MCC je verzija Pearsonovog koeficijenta korelacija koja je prilagođena radu s binarnim varijablama [73]. Oznake na slikama iste su kao one opisane kod prethodnih parametara - MAE i s. Simulacijske raspodjele prikazane su na *Slikama 3.13 i 3.14.*



Slika 3.16 Raspodjela koeficijenta korelacije (MCC) za varijable s omjerom klasa 50:50 %



Slika 3.17 Raspodjela koeficijenta korelacije (MCC) za varijable s omjerom klasa 80:20 %

Iz *Slika 3.16 i 3.17* vidljiv je utjecaj promjene udjela klase 1 na maksimalne vrijednosti i raspone raspodjela vrijednosti parametra MCC, dok su srednje vrijednosti raspodjela ostale nepromijenjene. Više numeričkih detalja dano je u *Tablici 3.7.*

Kod parametra MCC došlo je do suženja raspona apsolutnih vrijednosti pa je apsolutna razlika 1.25 za skup s udjelima 80:20 %, dok za skup 50:50 % raspon iznosi 2.0. Nešto manje odstupanje raspona je kod simulacija i ono iznosi 0.75 za skup 80:20 % i 0.88 za skup 50:50 %.

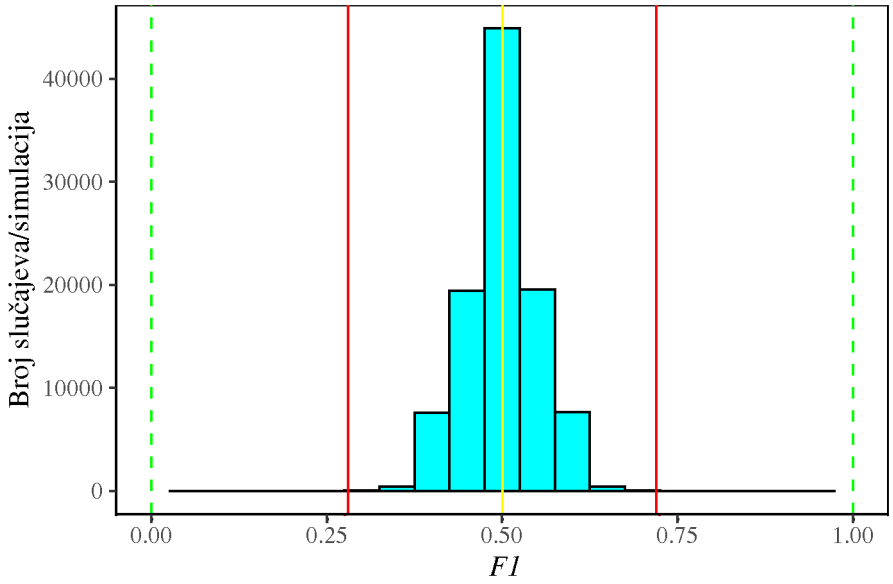
Tablica 3.7 Deskriptivni podaci raspodjela koeficijenta korelacije (MCC) za varijable s udjelima klasa 50:50 % i 80:20 %

	MCC_{50}	MCC_{80}
Minimum	-0.440	-0.250
Maksimum	0.440	0.500
Medijan	0.000	0.000
Srednja vrijednost	0.000	0.000
Apsolutni minimum	-1	-0.25
Apsolutni maksimum	1	1

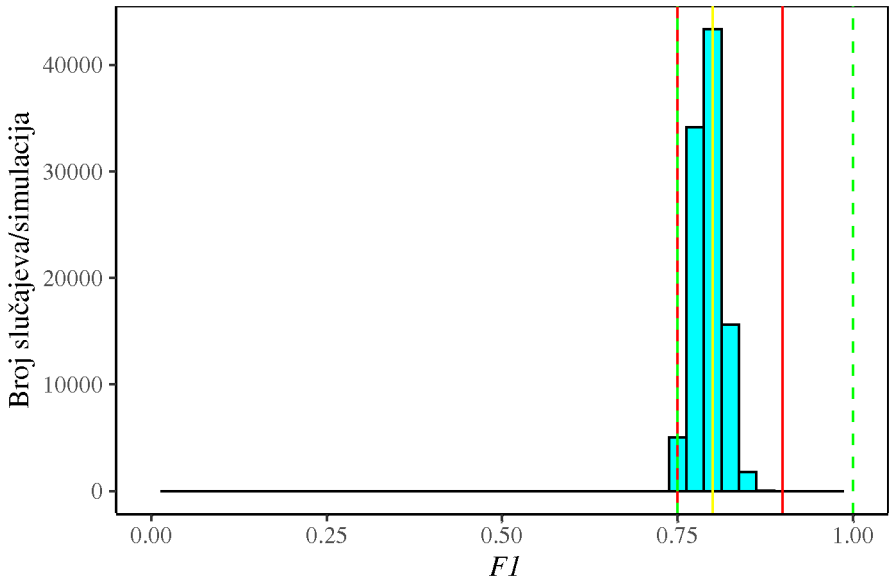
Kao i kod svih ostalih simulacija, vidljiva je promjena raspona na jednoj strani (u ovom slučaju lijevoj) gdje je došlo do velikog suženja i preklapanje minimalne apsolutne i minimalne simulacijske vrijednosti za skup 80%. Ova simulacija potvrđuje da promjena u omjeru klasa izaziva promjene u raspodjeli parametra MCC [74].

Raspodjela parametra *F1*

Parametar $F1$ je mjera preciznosti testova za rad s binarnim varijablama. Najniža vrijednost $F1$ je 0, a najviša 1 [37]. Prikaz simulacijskih raspodjela vrijednosti parametra $F1$ za dva udjela klasa dan je na *Slikama 3.18 i 3.19*. Oznake na slikama imaju analogno značenje kao kod parametara MAE ili MCC.



Slika 3.18 Raspodjela parametra $F1$ za varijable s omjerom klasa 50:50 %



Slika 3.19 Raspodjela parametra $F1$ za varijable s omjerom klasa 80:20 %

Iz *Slika 3.18 i 3.19* vidljiv je utjecaj promjene udjela klase 1 na promjenu raspona parametra $F1$ što su primijetili i drugi autori [75]. Više detalja o promjenama raspodjela, raspona i srednjih vrijednosti i medijana dostupno je u *Tablici 3.8.*

Tablica 3.8 Deskriptivni podaci raspodjele parametra $F1$ za varijable s udjelima klasa 50:50 % i 80:20 %

	$F1_{50}$	$F1_{80}$
Minimum	0.280	0.75
Maksimum	0.720	0.9
Medijan	0.5	0.8
Srednja vrijednost	0.5002	0.8
Apsolutni minimum	0	0.75
Apsolutni maksimum	1	1

Kod parametra $F1$ došlo je do suženja raspona apsolutnih vrijednosti pri čemu je apsolutna razlika 0.25 za skup 80:20 %, a 1.0 za skup 50:50 %. Nešto manja, ali i dalje znatna, promjena raspona simulacijskih vrijednosti iznosi 0.44 za skup 80:20 % i 0.15 za skup 50:50 %. Kao i kod svih ostalih simulacija, vidljiva je promjena raspona na jednoj strani (u ovom slučaju tamo gdje je $x \geq 50$ %) gdje je došlo do velikog suženja i preklapanje minimalne apsolutne i minimalne simulacijske vrijednosti za skup 80:20 %. Simulacija potvrđuje da promjena u omjera klasa utječe na promjenu u raspodjeli parametra $F1$ [75].

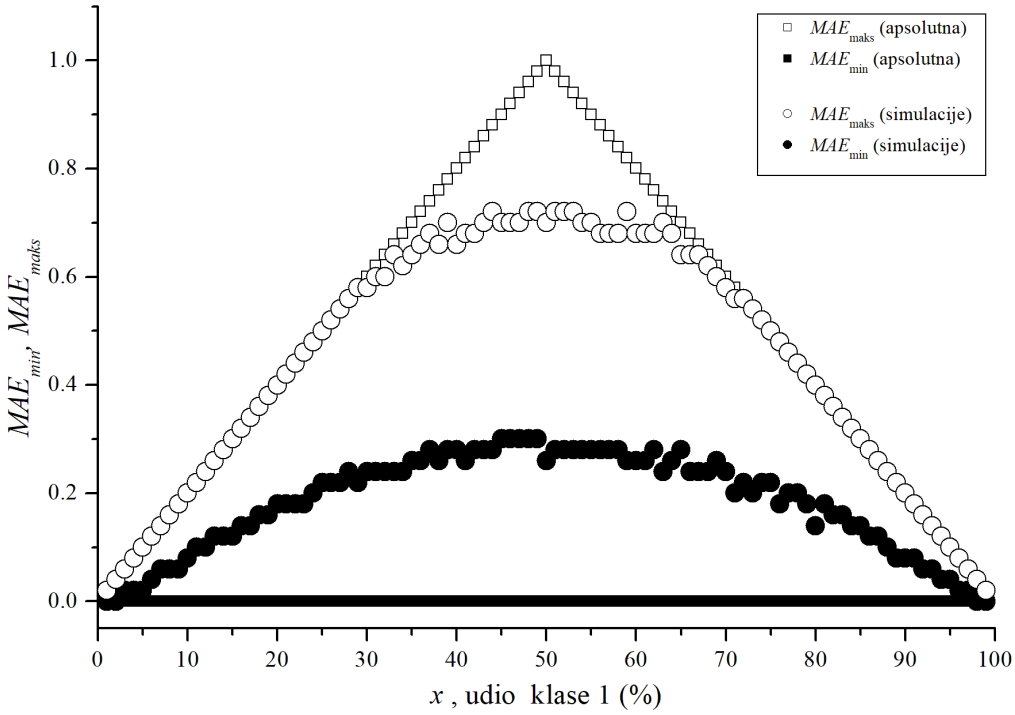
3.4.3 Usporedba simulacijskih i izvedenih karakterističnih vrijednosti parametra *MAE, s, MCC* i *F1*

U ovom dijelu provest će se usporedba izvedenih (poglavlje 3.4.1) i simulacijskih rezultata (poglavlje 3.4.2) dobivenih za karakteristične vrijednosti dodatnih parametara koji se koriste za izračun kvalitete modela, tj. parametara MAE, s, MCC i $F1$. Bit će provedena usporedba u širokom rasponu vrijednosti udjela klase 1 (x) u rasponu od 0.01 do 0.99 (od 1 % do 99 %). Rubne vrijednosti s udjelom klase 1 jednakim 0 % i 100% izostavljene su stoga što neki od parametara nisu definirani u tim slučajevima, te ih nije moguće jednostavno izračunati. Ova usporedba rezultata dobivenih izvodima i simulacijama, poslužit će i kao dodatna provjera i dokaz ispravnosti izvedenih matematičkih izraza za karakteristične vrijednosti parametara MAE, s, MCC i $F1$.

Rezultati simulacija grafički su prikazani za: (1) izvedene (pa time i apsolutne) i (2) simulacijske minimalne i maksimalne vrijednosti parametara, kao i za njihove raspone.

Usporedba za prosječnu apsolutnu pogrešku (*MAE*)

Na *Slici 3.20* prikazane su izvedene i simulacijske minimalne i maksimalne vrijednosti parametra MAE u ovisnosti o udjelu jedne klase (x) u rasponu od 1 % do 99 %.

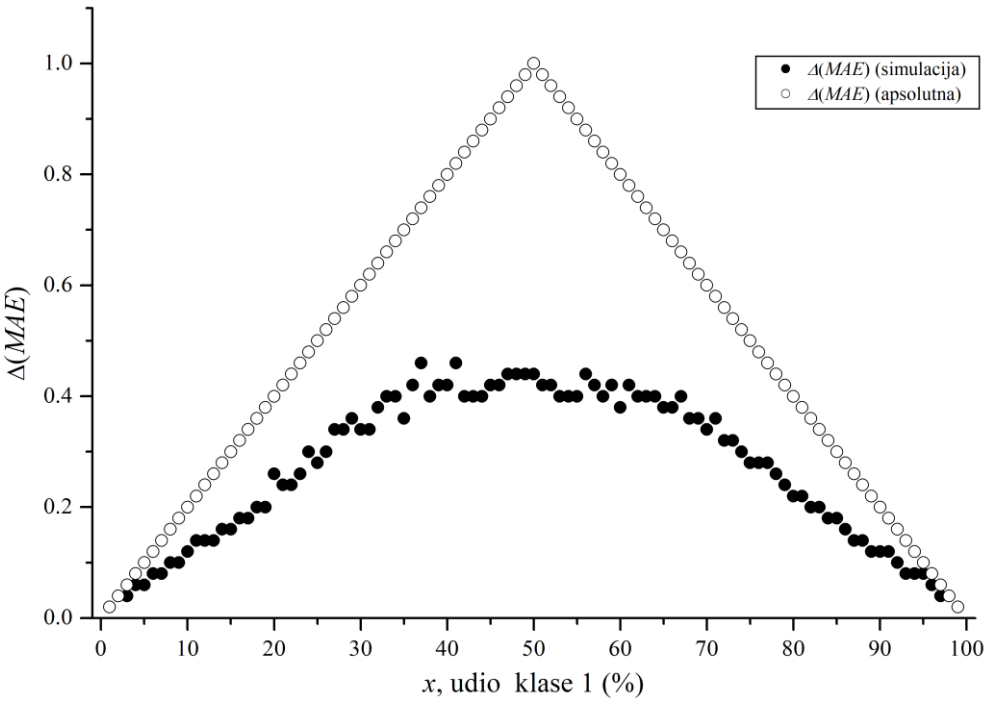


Slika 3.20 Apsolutne i simulacijske karakteristične vrijednosti MAE u ovisnosti o udjelu klase 1

Usporedbom parametra Q_2u grafu na *Slici 3.5* s MAE u grafu na slici 3.20., može se uočiti jasna međuovisnost ovih parametara jer minimum točnosti (Q_2) odgovara maksimalnoj pogrešci (MAE). Analogno vrijedi i za maksimalnu točnost, koja odgovara minimalnoj pogrešci. Uočava se da simulacijska vrijednost niti u jednom slučaju (za nijedan x) nije veća od maksimalne, niti manja od minimalne vrijednosti dobivene s pomoću izvedenih formula.

Kod parametra MAE vidi se da su apsolutne maksimalne vrijednosti parametra (izvedene iz formula) jednake simulacijskim za udjele klase 1 manje od 30% i veće od 70%. Slično kao kod parametara točnosti, do potpunog preklapanja nije došlo kod udjela bližim 50:50 % stoga što je broj mogućih permutacija modelne varijable M za te udjele daleko veći od 100.000, koliko je kreirano parova varijabli E i M u svakoj simulaciji (za odabrani x (%)).

Oduzimanjem minimalnih od maksimalnih vrijednosti parametra izračunan je raspon parametra MAE u ovisnosti o udjelu klase 1 (x) (*Slika 3.21*).



Slika 3.21 Apsolutni i simulacijski rasponi MAE u ovisnosti o udjelu klase 1

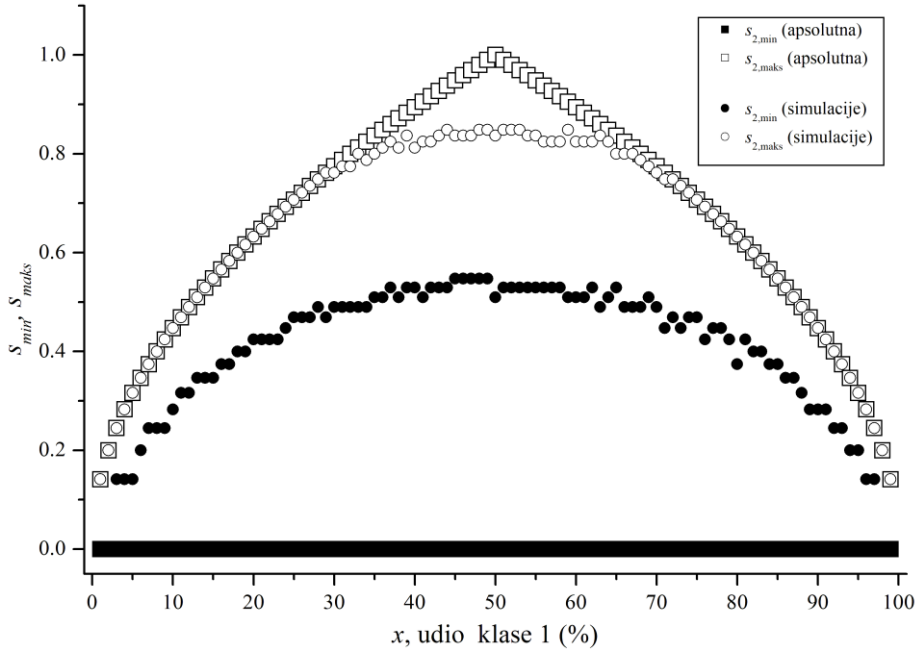
Apsolutna vrijednost raspona je potpuno istog oblika kao za parametar $\Delta\ Q_2$ (*Slika 3.8*), a identični su i matematički izrazi za raspon MAE (izraz (3.57)) i $\Delta\ Q_2$ (jednadžba (3.7)) u ovisnosti o udjelu klase 1 (x).

$$\Delta(MAE) = 1 - |1 - 2x|, \forall x \in [0,1] \tag{3.57}$$

Na slici se vidi da su sve simulacijske vrijednosti unutar raspona apsolutnih vrijednosti.

Karakteristične vrijednosti standardne pogreške – *s*

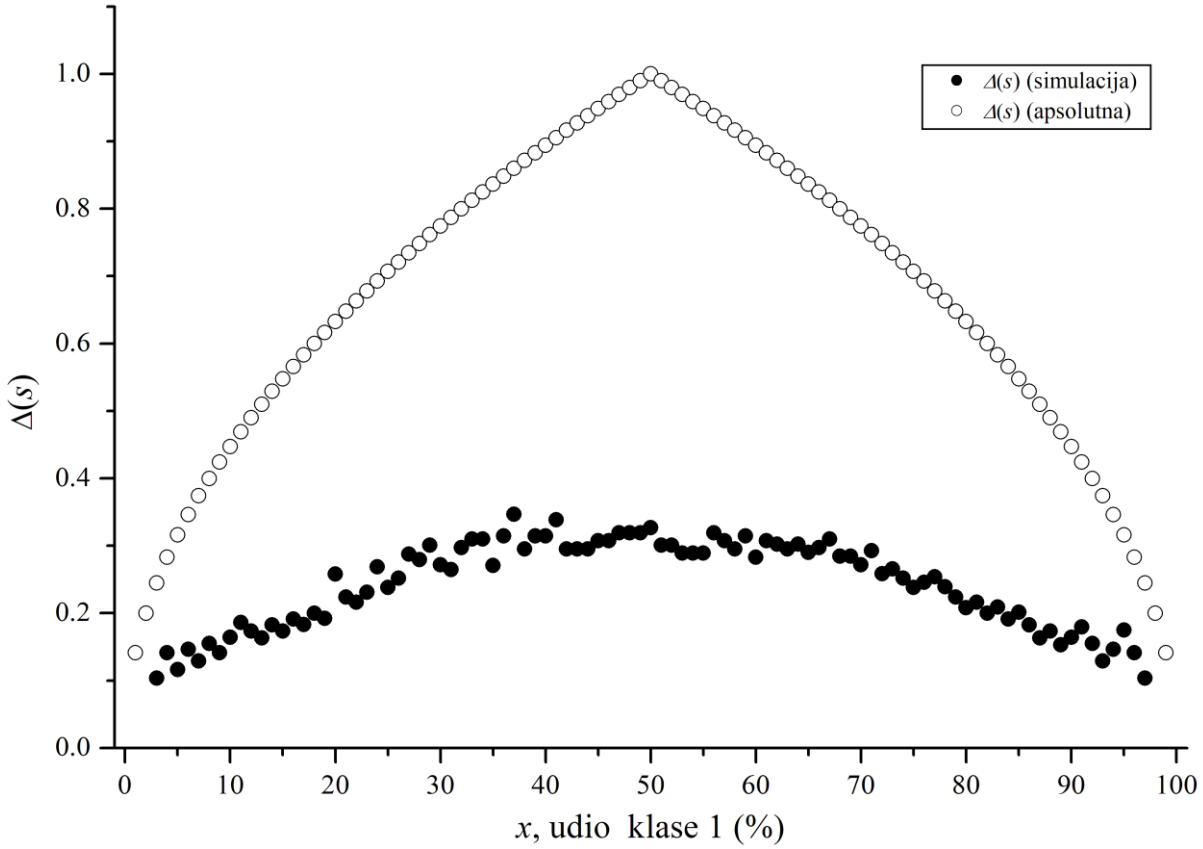
Simulacijske i izvedene minimalne i maksimalne vrijednosti standardne pogreške u ovisnosti o udjelu klase 1 (x), prikazane su na *Slici 3.22.*



Slika 3.22 Apsolutne i simulacijske karakteristične vrijednosti s u ovisnosti o udjelu klase 1

Ovisnost standardne pogreške o udjelu klase 1 (x) sličan je onom za ΔQ_2 sa *Slike 3.7*.

Razlika minimalne i maksimalne vrijednosti standardne pogreške daje raspon parametra, a prikazana je na *Slici 3.23* u ovisnosti o udjelu klase 1.



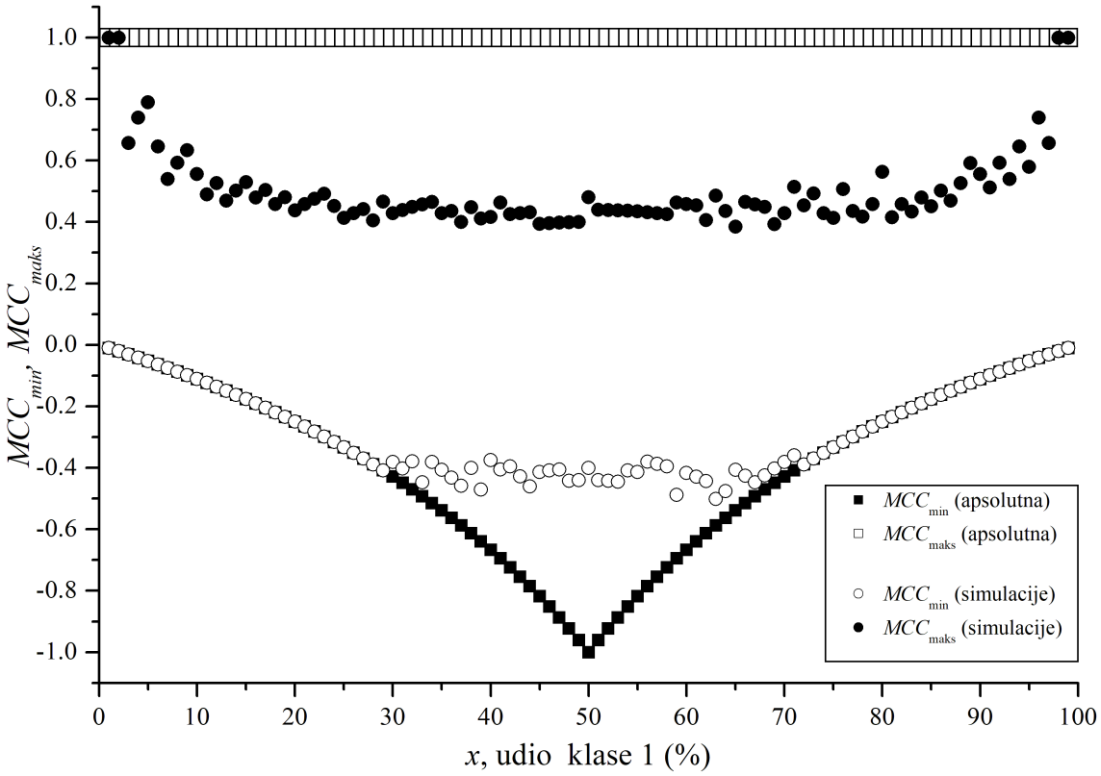
Slika 3.23 Apsolutni i simulacijski rasponi vrijednosti s u ovisnosti o udjelu klase 1

Formula za raspon parametra s u ovisnosti o udjelu klase 1 (x) dana je izrazom (3.58) i vrijedi na cijelom intervalu x.

$$\Delta(s) = \overline{1 - |1 - 2x|}, \forall x \in [0,1] \tag{3.58}$$

Karakteristične vrijednosti parametra *MCC*

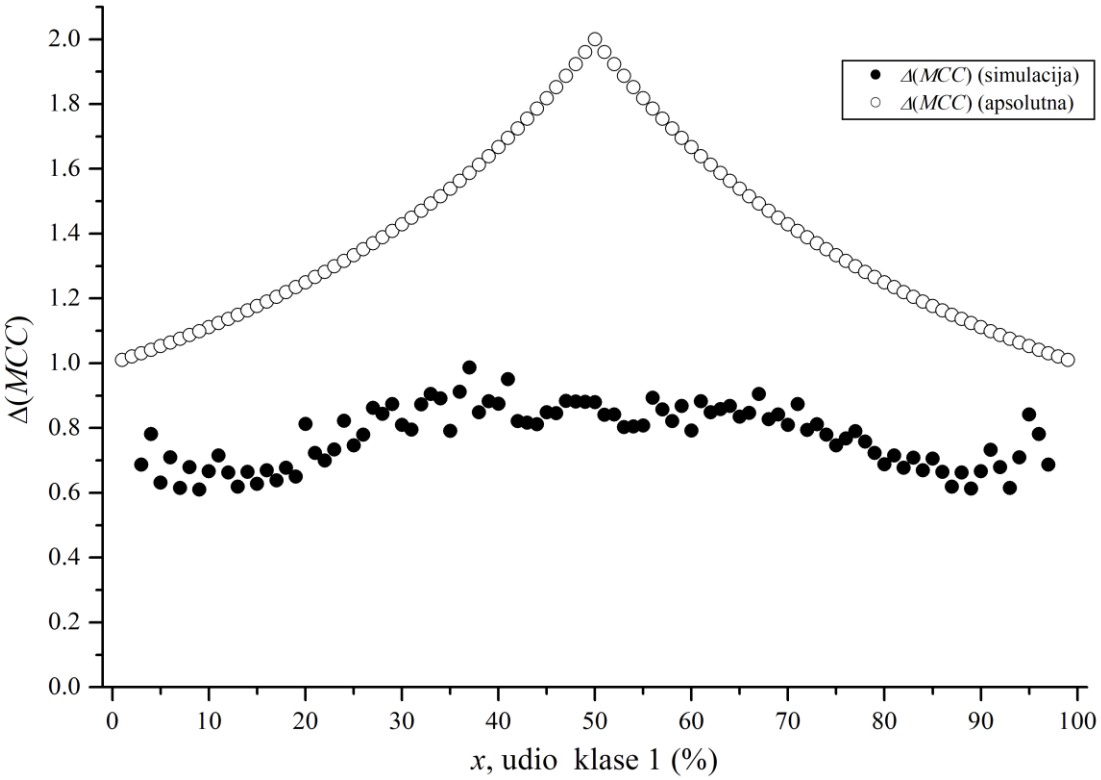
Simulacijske i apsolutne vrijednosti minimalne i maksimalne, vrijednosti parametra MCC prikazane su na grafu (*Slika 3.24*).



Slika 3.24 Apsolutne i simulacijske karakteristične vrijednosti MCC u ovisnosti o udjelu klase 1

Na grafu (*Slika 3.24*) vidimo da su sve simulacijske vrijednosti ili jednake (za udjele klase 1 ispod 30 % i iznad 70 %, kao i kod većine drugih parametara) ili veće od apsolutnih minimalnih vrijednosti parametra MCC dobivenim formulama (3.46) i (3.47) za lijevi i desni pod-interval udjela klase 1. Za udio klase 1 koji je bliže 50%, broj pretraženih nasumičnih permutacija varijable M - koji uvijek iznosi 100.000, premali je u odnosu na ukupni mogući broj permutacija (~ 10^{29}) za varijablu koja ima ukupno 100 vrijednosti. Kao i kod parametra točnosti Q_2, maksimalna apsolutna vrijednost parametra MCC u ovisnosti o udjelu klase 1 (x) konstantna je i jednaka 1.0. Tomu je razlog činjenica da varijable E i M imaju identične udjele klase 1 i 0, pa zbog toga sigurno postoji permutacija varijable M koja će biti identična redoslijedu vrijednosti varijable E. U tom slučaju će se te dvije varijable savršeno podudarati. Izvodi karakterističnih vrijednosti parametra MCC dani su u prilozima (*Prilozi 3.24 -3.29*).

Raspon parametra MCC prikazan je na grafu (*Slici 3.25*).

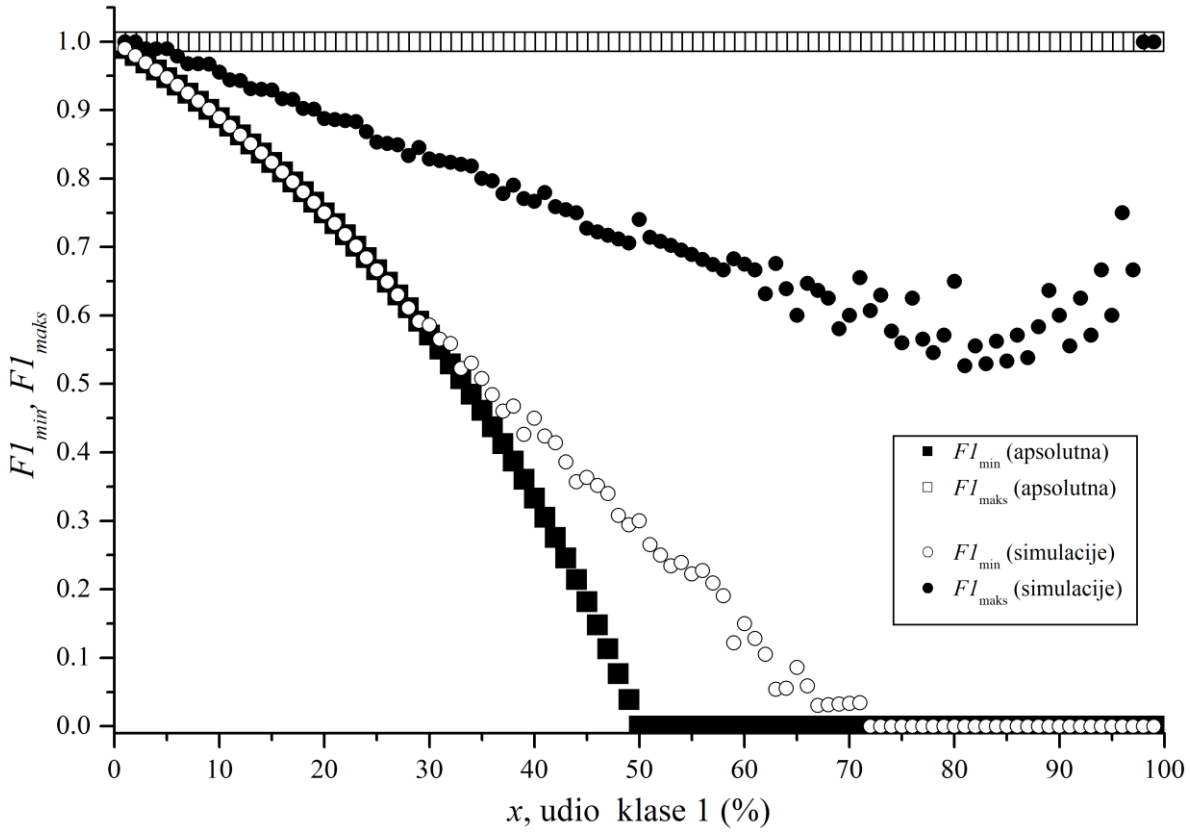


Slika 3.25 Apsolutni i simulacijski rasponi vrijednosti parametra MCC u ovisnosti o udjelu klase 1

Vrijednosti prikazane na *Slici 3.25* potvrđuju da se rasponi dobiveni simulacijama nalaze u okvirima apsolutnih raspona parametra MCC, što posredno ukazuje na ispravnost izvedenih izraza za raspon MCC.

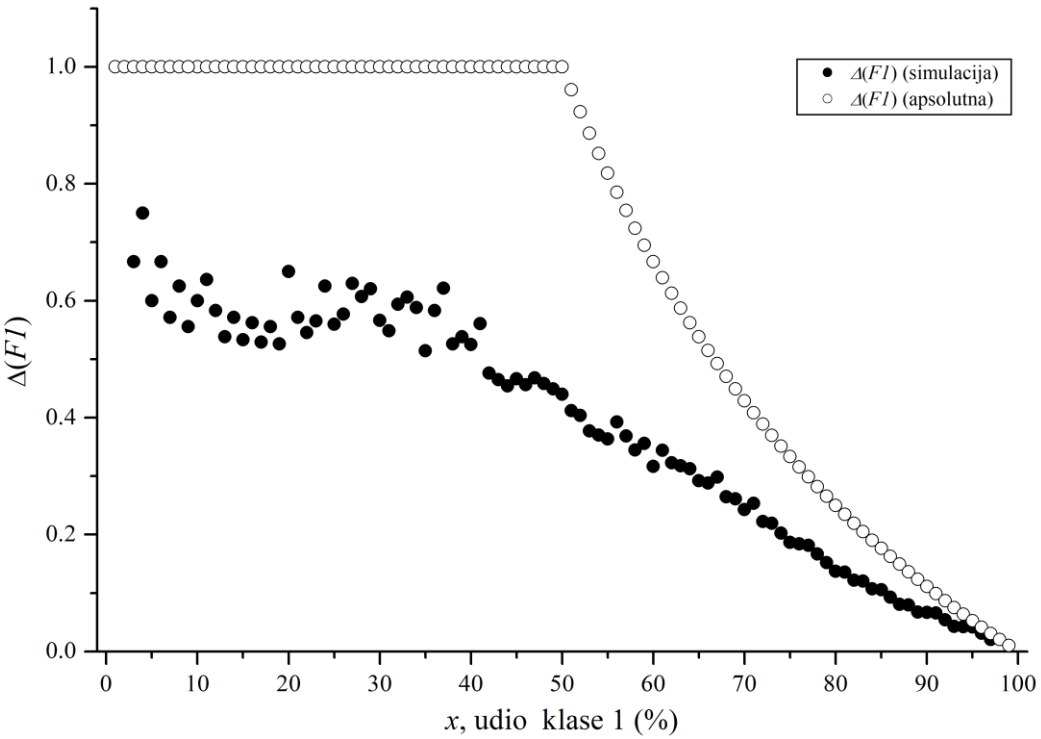
Karakteristične vrijednosti parametra $F1$

Za razliku od svih drugih parametara, parametar $F1$ asimetričan je u odnosu na udio klasa 50:50 %, što znači da obična zamjena oznaka klase 1 i 0 može uzrokovati potpuno različiti rezultat. Ovisnost parametra $F1$ o udjelima klase 1, prikazana je na grafu (*Slika 3.26*).



Slika 3.26 Apsolutne i simulacijske karakteristične vrijednosti $F1$ u ovisnosti o udjelu klase 1

Ponovno se mogu primijetiti poklapanje simulacijskih vrijednosti s apsolutnim za udjele manje od 30 % i veće od 70 %, dok kod maksimalnih vrijednosti poklapanja nema, a razlog je opisan kod parametra srednje apsolutne pogreške MAE (u tekstu ispod *Slike 3.20*) i koeficijenta korelacije MCC (u tekstu ispod *Slike 3.24*). Sve simulacijske vrijednosti nalaze se između minimalnih i maksimalnih apsolutnih vrijednosti, a raspon vrijednosti parametra $F1$ prikazan je na *Slici 3.27*.



Slika 3.27 Apsolutni i simulacijski rasponi parametra $F1$ u ovisnosti o udjelu klase 1

Ovisnost raspona o udjelu klase 1 pokazuje sličnu asimetriju kao i vrijednosti na grafu (*Slika 3.26*).

Nedostatak parametra $F1$ njegova je asimetrija uslijed favoriziranje jedne od dvije klase, što rezultira značajno većim rasponom vrijednosti za manje vrijednosti udjela klase 1. To može izazvati problem neispravnog rangiranja predviđanja, tj. prediktivnih metoda/modela.

3.5 Usporedba izvedenih karakterističnih vrijednosti parametara MAE, s, MCC i $F1$ s entropijom

U *Tablici 3.9* dani su koeficijenti korelacije u rasponu udjela klase 1 od 1 % do 50 % (i od 0 % do 50%) između entropije varijabli ($logW$), s jedne strane, te karakterističnih vrijednosti parametara MAE, s, MCC i $F1$ i njihovih razlika s druge strane. Radi usporedbe s ranijim rezultatima za parametar točnosti Q_2, korelacije za taj parametar dane su u zadnjem retku *Tablice 3.9 A*) i *B*). S obzirom na simetričnu ovisnost $logW$ i parametara kvalitete o udjelu klase 1 (x) u odnosu na vrijednost x = 50 %, korelacije su računane samo do tog udjela. Nadalje, neki parametri nisu dobro definirani za x = 0 %, pa su korelacije računane polazeći od udjela x = 0 % (*Tablica 3.9, A*), kao i od udjela x = 1 % (*Tablica 3.9, B*).

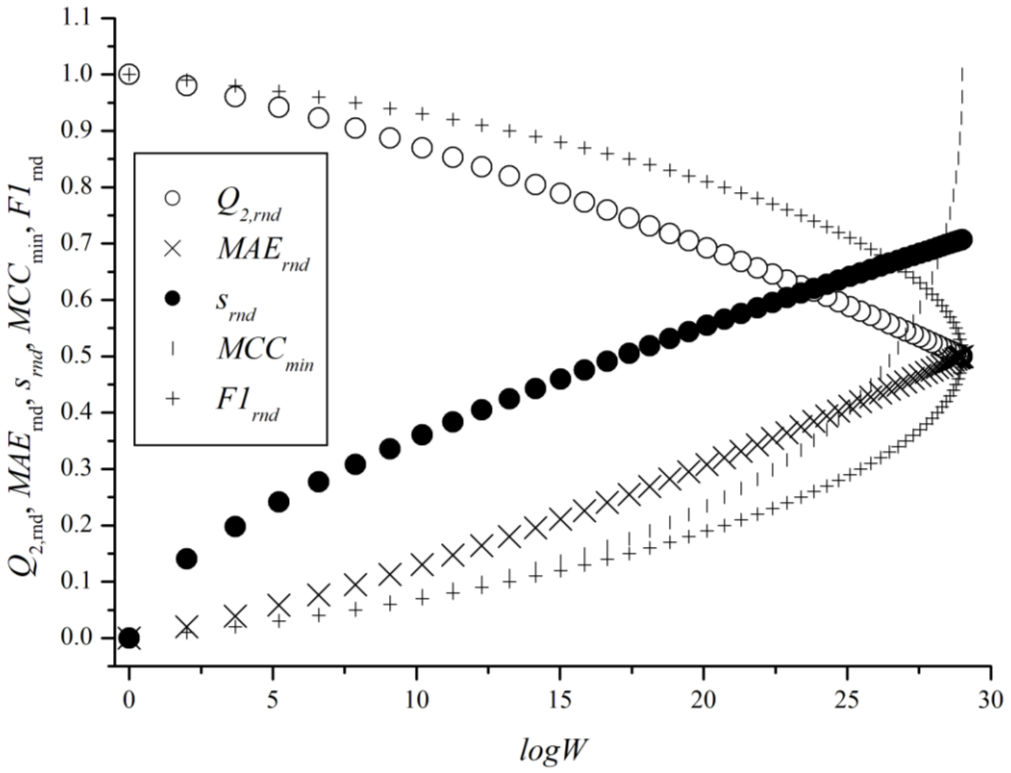
Tablica 3.9 Koeficijenti korelacije između entropije ($logW$) i apsolutnih minimalnih i maksimalnih vrijednosti parametara za varijable u rasponu udjela klase 1 (x) od 1 % do 50 % (A) i od 0 do 50 %.

parametar	karakteristične vrijednosti		apsolutni rasponi	
	rubne[a]	prosječne nasum.	maks. − prosj.	min. − prosj.
A) Koeficijenti korelacije za varijable s udjelom klase 1 (x) od 1 % do 50 %				
MAE	0.948	0.997	0.842	0.997
s	0.989	0.990	0.880	0.997
MCC^b	-0.879	-[c]		0.879
$F1^b$	-	0.948		0.948
Q_2	-0.948	-0.997	0.997 (ΔQ_2)	0.842
B) Koeficijenti korelacije za varijable s udjelom klase 1 (x) od 0 % do 50 %				
MAE	0.946	0.996	0.833	0.996
s	0.990	0.990	0.873	0.990
MCC^b	-0.875	-		0.875
$F1$	-	0.946		0.946
Q_2	-0.946	-0.996	0.996 (ΔQ_2)	0.833

[a]Rubne karakteristične vrijednosti su za pogreške (MAE i s) maksimalne (minimalne su konst. = 0), a za točnost/podudarnost (MCC i $F1$) minimalne (maksimalne su konst. = 1) vrijednosti; [b]Maksimalne vrijednosti parametara MCC i $F1$ jednake su 1, a prosječna nasumična vrijednost $MCC = 0$; [c]Prosječna nasumična vrijednost je konstantna, pa nije moguće izračunati koeficijent korelacije.

Najveću korelaciju s $logW$ ($R = 0.997$) u cijelom rasponu klase 1 (x) od 0 % do 50 % imaju prosječne nasumične vrijednosti parametara MAE i Q_2, odnosno parametar stvarnog doprinosa modela iznad nasumične točnosti ΔQ_2 (*Tablica 3.9*, dio *B)*. Pritom, sam predznak korelacije je logičan jer što je entropija varijabli E i M veća, to je veće odstupanje između. Odmah nakon ta dva parametra, sljedeća najbolja korelacija u istom rasponu je s parametrom standardne pogreške s ($R = 0.990$). Taj je zaključan logičan jer se pogreške MAE i s računaju na temelju elemenata tablice pogrešaka u i o (gdje je u slučaju izmjenjivih varijabli $u = o$), dok se točnost računa simetrično - na temelju točnih pogađanja klase 1 (p) i 0 (n).

Grafički prikaz ovisnosti između parametra $logW$ koji predstavlja entropiju varijable i karakterističnih vrijednosti parametara Q_2, MAE, s, MCC i $F1$ prikazan je na *Slici 3.28*.



Slika 3.28 Grafički prikaz ovisnost entropije ($logW$) o karakterističnim vrijednostima parametara Q_2, MAE, s, MCC i $F1$ koje pokazuju najbolje slaganje

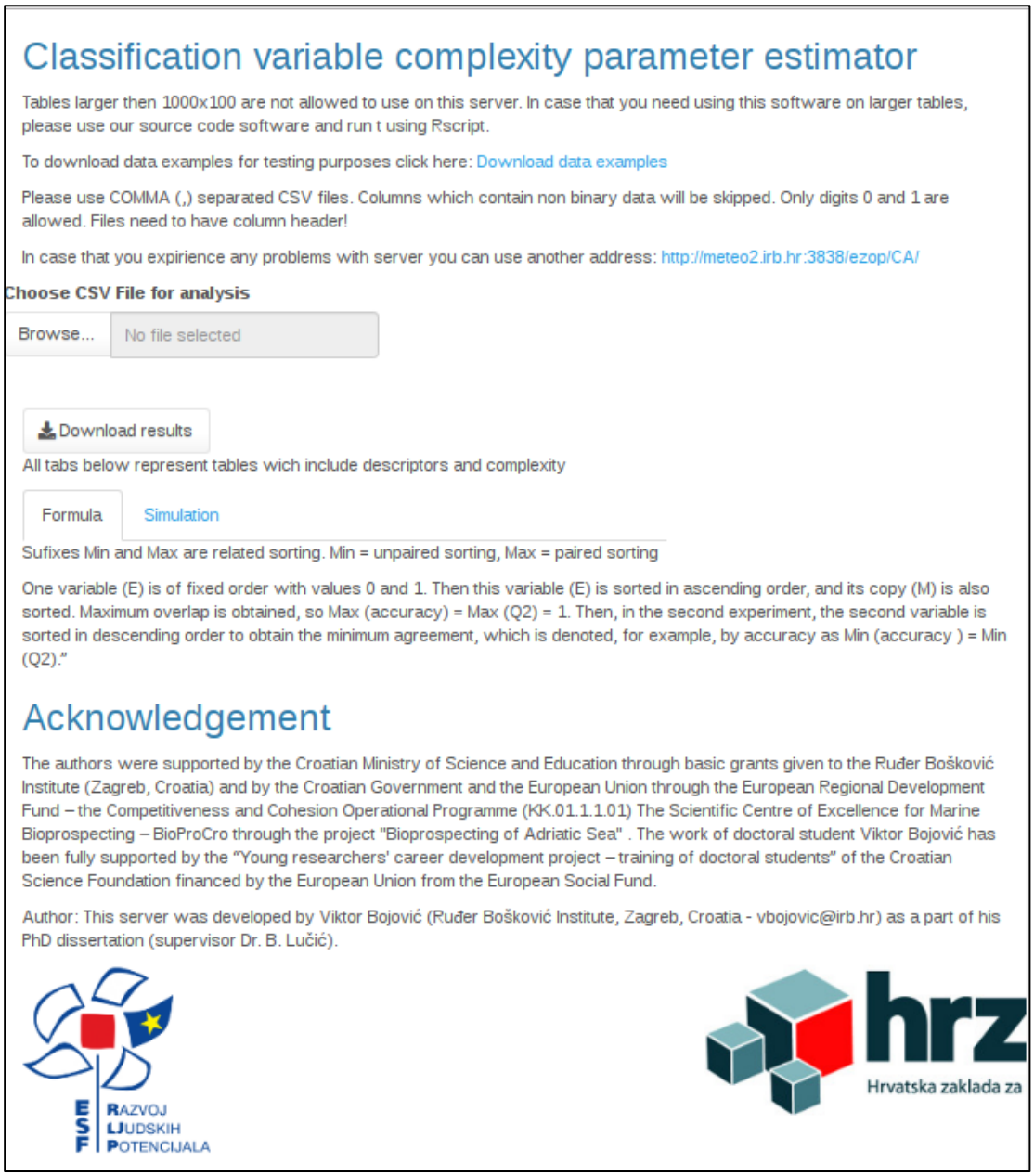
Jasno se može razaznati sličnost ovisnosti o entropiji parametra za $Q_{2,rnd}$ s onim za MAE_{rnd} i s_{rnd}, i ta je ovisnost linearna. Ovdje se u slučaju uravnoteženih modela odnosno izmjenjivih varijabli pokazuje potpuna identičnost ovisnosti entropije varijable $logW$ i $Q_{2,rnd}$ (ili $\Delta Q_2 = Q_{2,max} - Q_{2,rnd} = 1 - Q_{2,rnd}$) s ovisnošću $logW$ i MAE_{rnd} (odnosno $MAE_{min} - MAE_{rnd} = 0 - MAE_{rnd} = -MAE_{rnd}$). Ovo je i razumljivo, jer dok $1 - Q_{2,rnd}$ daje maksimalni mogući stvarni doprinos modela mjereći točnost/podudarnost između varijabli E i M koja je iznad nasumične, MAE_{rnd} u tom slučaju točnost/podudarnost iskazuje preko $u + o$ (odnosno preko $u + o = 2u$, s obzirom da je kod izmjenjivih varijabli $o = u$), tj. minimalne moguće srednje apsolutne pogreške/nepodudarnosti. Dakle, podudarnost između varijabli E i M iskazuje se preko zbroja točnih predviđanja $p + n$, a kako je ukupan broj vrijednosti u varijablama E i M jednak $N = (p + n) + (u + o) = (p + n) + (2u)$, jasno je da je to ekvivalentno mjerenju nepodudarnosti/pogreške preko $(2u)$. I jedan i drugi parametar može se iskazati u postotcima, što ističe ova dva parametra kao intuitivno jasna i lako razumljiva, te time i vrlo prikladna za uporabu. Ujedno, to su parametri koji pokazuju najbolju korelaciju s entropijom varijabli od svih isprobanih parametara i raspona u *Tablici 3.9.* Funkcionalna ovisnost najboljih karakterističnih vrijednosti koeficijenta korelacije (MCC_{min}) i parametra $F1$ ($F1_{min}$) o entropiji varijable ($logW$) izrazito je nelinearna. To zapravo upućuje na izvjesne slabosti tih parametara, što je primijećeno i ranije, a uočljivo je osobito za parametar $F1$ (*Slike 3.26* i *3.27*) koji zanemaruje točno predviđanje klase 0 kad je ona većinska klasa.

3.6 Primjena izvedenih parametara složenosti varijabli na podatcima u QSAR modeliranju

3.6.1 Izrada mrežnog poslužitelja za procjenu složenosti varijabli

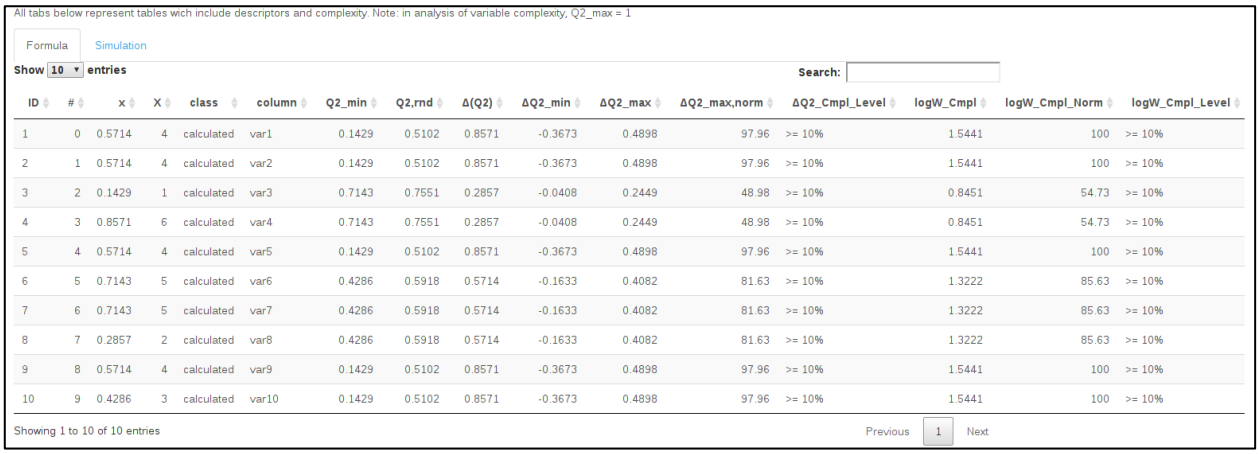
Classification variable complexity parameter estimator mrežna je aplikacija napravljena u svrhu izračuna i analize složenosti podataka, a funkcionira kao besplatni server (poslužitelj). Varijable (molekularni deskriptori) učitavaju se u obliku stupaca podatkovne tablice. Osnovni oblik varijabli je binarni, gdje su jedine moguće vrijednosti varijable 1 ili 0. Kako su obično stvarni skupovi deskriptora kombinacija binarnih, općenitih klasifikacijskih varijabli s više klasa i kontinuiranih varijabli, za lakši rad servera, na svim varijablama koje nisu čisto binarne provodi se postupak dihotomizacije. Taj je postupak opisan u nastavku.

Složenost varijabli i ostali karakteristični parametri (minimalne, maksimalne i prosječne nasumične vrijednosti) računaju se za sve varijable iz ulazne datoteke. Opis načina rada aplikacije i dijelova koda nalazi se u prilozima (*Prilog 3.44*). Web aplikacija dostupna je online [76], a njezin osnovni izgled prikazan je na *Slici 3.29*. Rezultati za svaku varijablu prikazuju se u recima, a oznake (značenja kratica) parametara koji se računaju na serveru, opisane su u nastavku. Primjeri prihvatljivih ulaznih datoteka dani su na mrežnoj stranici aplikacije.



Slika 3.29 Izgled mrežne aplikacije „*Classification variable complexity parameter estimator*"

Rezultati izračuna složenosti prikazani su u tabličnom obliku, i moguće je njihovo preuzimanje klikom na ikonu „*Download results*". Primjer tabličnog rezultata prikazan je na *Slici 3.30.*



All tabs below represent tables wich include descriptors and complexity. Note: in analysis of variable complexity, Q2_max = 1

Formula | Simulation

Show 10 entries | Search:

ID	#	x	X	class	column	Q2_min	Q2,rnd	Δ(Q2)	ΔQ2_min	ΔQ2_max	ΔQ2_max,norm	ΔQ2_Cmpl_Level	logW_Cmpl	logW_Cmpl_Norm	logW_Cmpl_Level
1	0	0.5714	4	calculated	var1	0.1429	0.5102	0.8571	-0.3673	0.4898	97.96	>= 10%	1.5441	100	>= 10%
2	1	0.5714	4	calculated	var2	0.1429	0.5102	0.8571	-0.3673	0.4898	97.96	>= 10%	1.5441	100	>= 10%
3	2	0.1429	1	calculated	var3	0.7143	0.7551	0.2857	-0.0408	0.2449	48.98	>= 10%	0.8451	54.73	>= 10%
4	3	0.8571	6	calculated	var4	0.7143	0.7551	0.2857	-0.0408	0.2449	48.98	>= 10%	0.8451	54.73	>= 10%
5	4	0.5714	4	calculated	var5	0.1429	0.5102	0.8571	-0.3673	0.4898	97.96	>= 10%	1.5441	100	>= 10%
6	5	0.7143	5	calculated	var6	0.4286	0.5918	0.5714	-0.1633	0.4082	81.63	>= 10%	1.3222	85.63	>= 10%
7	6	0.7143	5	calculated	var7	0.4286	0.5918	0.5714	-0.1633	0.4082	81.63	>= 10%	1.3222	85.63	>= 10%
8	7	0.2857	2	calculated	var8	0.4286	0.5918	0.5714	-0.1633	0.4082	81.63	>= 10%	1.3222	85.63	>= 10%
9	8	0.5714	4	calculated	var9	0.1429	0.5102	0.8571	-0.3673	0.4898	97.96	>= 10%	1.5441	100	>= 10%
10	9	0.4286	3	calculated	var10	0.1429	0.5102	0.8571	-0.3673	0.4898	97.96	>= 10%	1.5441	100	>= 10%

Showing 1 to 10 of 10 entries | Previous | 1 | Next

Slika 3.30 Prikaz rezultata mrežnog poslužitelja (servera) za izračun složenosti varijabli

U primjeru tablice na mrežnom poslužitelju prikazane su karakteristične vrijednosti parametara Q_2, $Q_{2,rnd}$, ΔQ_2 i MCC, te na kraju procjena složenosti varijable (parametar $logW$).

Prvi dio slike uključuje slijedeće stupce:

- ID – automatski dodijeljen redni broj retka
- # - redni broj stupca u ulaznoj datoteci
- *x* – udio klase 1
- *X* – broj elemenata kase 1 u varijabli
- class – oznaka načina izračuna parametara ('*calculated*' znači izračun formulama, a „*simulation*“ znači izračun pomoću sortiranja podataka)
- column – naziv varijable/stupca iz ulazne datoteke
- Q2_min – minimalna vrijednost parametra Q_2
- Δ(Q2) – raspon parametra Q_2
- ΔQ2_min – minimalna vrijednost parametra ΔQ_2
- ΔQ2_max – stvarni maksimalni doprinos modela - razlika parametara: $\Delta Q_2 = Q_{2,max} - Q_{2,rnd} = \mathrm{Q2_{max}} - \mathrm{Q2rnd}$
- Q2rnd - parametar $Q_{2,rnd}$ (prosječna nasumična vrijednost parametra Q_2)
- ΔQ2_max,norm – normalizirana vrijednost parametra složenosti ΔQ_2 ($\Delta Q_{2,max,norm}$)
- ΔQ2_Cmpl_Level – razina složenosti definiran pomoću maksimalne vrijednosti ΔQ_2 parametra najsloženije moguće varijable s jednakim brojem elemenata obje klase (po $N/2$) ili najbliže moguće tom broju, kod varijabli s neparnim brojem vrijednosti (N)
- logW_Cmpl – složenost ($logW$)
- logW_Cmpl_Norm – normalizirana složenost/entropija varijable $(logW)_{norm}$
- logW_Cmpl_Level – razina složenosti definirana parametrom $logW$

3.6.2 Primjena rezultata u analizi skupova varijabli iz literature

Kontinuirane varijable posebne su po tome što između svake vrijednosti može postojati beskonačan broj vrijednosti, ukoliko je to tehnički moguće. Kako bi rezultati ovog rada bili primjenjivi na kontinuiranim varijablama, potrebno ih je prilagoditi - transformacijom vrijednosti u dvije klase. Ta

se pretvorba provodi tako da se svaka vrijednost uspoređuje sa srednjom vrijednošću varijable. Vrijednosti manje od srednje vrijednosti svrstavaju se u klasu „0“, a one veće ili jednake srednjoj vrijednosti u klasu „1“.

Za analizu rada aplikacije sa stvarnim podacima koristit će se nekoliko skupova deskriptora od kojih su neki podskupovi binarne klasifikacijske, neki općenite klasifikacijske a neki kontinuirane varijable. Stoga će, radi jednostavnosti provedbe analiza, svi podaci biti dihotomizirani u odnosu na srednju vrijednost [66,67]. Korisniku aplikacije *Classification variable complexity parameter estimator* [76] bit će dana informacija o razini složenosti pojedine varijable iz skupa deskriptora koji su predani za analizu. Pritom, bit će izračunane i dane dvije razine složenosti - gledano prema normaliziranoj entropiji $logW$ ($logW_{norm}$) i prema normaliziranom maksimalnom doprinosu modela ΔQ_2 ($\Delta Q_{2,max,norm}$).

Maksimalna vrijednost parametra ΔQ_2 računa se po formuli $\Delta Q_2 = 1 - Q_{2,rnd}$ (*Prilog 3.11*) i bit će normalizirana kako bi joj vrijednost bila između 0 i 1. Maksimalna vrijednost doprinosa modela ΔQ_2 za binarne varijable računa se pomoću formule $Q_{2,rnd} = 2x^2 - 2x + 1$ (*Prilog 3.8*), pa je $\Delta Q_2 = -2x^2 + 2x$. Kako je najveći mogući stvarni doprinos točnosti neke varijable iznad nasumične točnosti ΔQ_2 jednak 0.5, ta je vrijednost korištena u *Tablici 3.10* kao faktor normalizacije u nazivniku.

Maksimalna vrijednost parametra ΔQ_2 računa se po formuli $\Delta Q_2 = 1 - Q_{2,rnd}$ (*Prilog 3.11*) i bit će normalizirana kako bi joj vrijednost bila između 0 i 1. Maksimalna vrijednost doprinosa modela ΔQ_2 za binarne varijable računa se pomoću $Q_{2,rnd} = 2x^2 - 2x + 1$ (*Prilog 3.8*), pa je rezultat $\Delta Q_{2,max} = -2x^2 + 2x$. Kako je najveći mogući stvarni doprinos točnosti neke varijable iznad nasumične točnosti ΔQ_2 jednak 0.5, ta je vrijednost korištena u *Tablici 3.10* kao faktor normalizacije u nazivniku.

Analiza skupa s 30 deskriptora iz rada Huuskonen [66] (*Huuskonen30*, *Prilog E_3.6*) aplikacijom *Classification variable complexity parameter estimator* [76] pokazala je da neki deskriptori nemaju složenost veću od 10 %. Skup je izvorno priređen za QSPR modeliranje topljivosti organskih spojeva u vodi s pomoću multivarijatne regresije i neuronskih mreža. Oni deskriptori čija je razina složenosti prema $\Delta Q_{2,max,norm}$ parametru manja od 10 % prikazani su u *Tablici 3.10*. Deskriptori imaju po 884 vrijednosti (za jednako toliko molekula u skupu), pa se taj skup može smatrati velikim skupom.

Tablica 3.10 Deskriptori iz skupa *Huuskonen30* najniže složenosti prema normaliziranim parametrima $\Delta Q_{2,norm}$ i $logW_{norm}$

deskriptor	X	x (%)	$\Delta Q_{2,max,norm}$	prag složenosti $\Delta Q_{2,max,norm}$	$logW_{norm}$	prag složenosti $logW_{norm}$
SssssN+	2	0.23	0.9 %	< 1 %	2.11 %	≥ 1 %
SsSH	8	0.9	3.59 %	≥ 2.5 %	7.16 %	≥ 5 %
SdS	16	1.8	7.11 %	≥ 5 %	12.76 %	≥ 10 %
SdssS	881	99.6	1.35 %	≥ 1 %	3.05 %	≥ 2.5 %
SsI	16	1.8	7.11 %	≥ 5 %	12.76 %	≥ 10 %

U *Tablici 3.10* stupac „*X*“ označava broj jedinica (broj elemenata u skupu koji pripadaju klasi 1) dobiven nakon dihotomizacije varijabli pomoću srednje vrijednosti. Na primjer, deskriptor

„SssssN+“ ima samo dvije vrijednosti koje nisu nula - od njih ukupno 884. Stupcem „x“ označen udio klase 1 u varijabli s ukupno vrijednosti. Brojem 1 označena je klasa koja je veća ili jednaka srednjoj vrijednosti. Stupac $\Delta Q_{2,max,norm}$ označava normiranu vrijednost parametra ΔQ_2 izračunanu pomoću udjela (x) te normiranu u odnosu na najsloženiju varijablu s N vrijednosti od kojih je polovica u jednoj klasi. Analogno, isto je pravilo normalizacije primijenjeno na entropiju varijable $logW$ (Boltzmannovu entropiju), te je dobivena normalizirana vrijednost $logW_{norm}$ koja se koristi kao mjera složenosti.

Iz *Tablice 3.10* vidimo da je niska složenost onih deskriptora koji imaju većinu vrijednosti u jednoj klasi. Deskriptori SssssN+ i SdssS svakako nemaju dovoljnu složenost za uključivanje u bilo kakav model. Također, slično je i s deskriptorom SsSH, pa i s deskriptorima SdS i SsI. Međutim, važno je napomenuti da su u modelima u [66] svi ovi deskriptori bili uključeni u model.

Slična obrada provedena je i na proširenom skupu koji sadrži 58 deskriptora (*Huuskonen58*), svaki s 884 vrijednosti [66], a rezultati su prikazani u *Tablici 3.11*.

Tablica 3.11 Značajnost deskriptora iz skupa *Huuskonen58* prema normaliziranim parametrima $\Delta Q_{2,norm}$ i $logW_{norm}$

deskriptor	$\Delta Q_{2,max,norm}$	prag složenosti $\Delta Q_{2,max,norm}$	$logW_{norm}$	prag složenosti $logW_{norm}$
StCH	2.7 %	≥ 2.5 %	5.6 %	≥ 5 %
SddC	0.45 %	< 1 %	1.11 %	≥ 1 %
StsC	7.98 %	≥ 7.5 %	14.04 %	≥ 10 %
SdNH	0.9 %	< 1 %	2.11 %	≥ 1 %
SaaNH	0.9 %	< 1 %	2.11 %	≥ 1 %
StN	4.92 %	≥ 2.5 %	9.37 %	≥ 7.5 %
SaasN	0.45 %	< 1 %	1.11 %	≥ 1 %
SssssN+	0.9 %	< 1 %	2.11 %	≥ 1 %
SaaO	5.36 %	≥ 5 %	10.07 %	≥ 10 %
SdsssP	4.03 %	≥ 2.5 %	7.92 %	≥ 7.5 %
SsSH	3.59 %	≥ 2.5 %	7.16 %	≥ 5 %
SdS	7.11 %	≥ 5 %	12.76 %	≥ 10 %
SaaS	5.36 %	≥ 5 %	10.07 %	≥ 10 %
SdssS	1.35 %	≥ 1 %	3.05 %	≥ 2.5 %
SsI	7.11 %	≥ 5 %	12.76 %	≥ 10 %

U *Tablici 3.11* najmanju značajnost prema ΔQ_2 parametru imaju deskriptori SaasN i SddC gdje samo jedna od 884 vrijednosti nije jednaka 0. U slučaju deskriptora SdNH, SaaNH i SssssN+ samo dvije od 884 vrijednosti nisu jednake 0.

Za oba skupa podataka (*Huuskonen30* i *Huuskonen58*) pokazalo se da samo malen broj deskriptora ne odgovara razini složenosti dovoljnoj za obradu. Međutim, nedostatak je modela iz rada [66] koji su temeljeni na svim deskriptorima taj što nisu prepoznati i isključeni iz modela deskriptori vrlo niske (tj. zanemarive) složenosti.

Treći skup deskriptora koji će biti analiziran sadrži 33 analoga taksana. Skup sadrži samo tri deskriptora od kojih je jedan indikatorski deskriptor (I_{HAL}) (objašnjen u uvodu disertacije).

Indikatorski deskriptor predstavljen je u radu Verme i Hanscha [67] (Tablica 4 iz toga rada) za potrebe modeliranje inhibicije rasta stanica raka log(1/IC50) analozima taksana (*Tablica 3.12)*.

Tablica 3.12 Skup deskriptora i biološke aktivnosti 33 spoja iz skupine taksana (*taksani*) u QSAR modeliranju preuzeto iz rada Verme i Hanscha [67]

No.	R	X	Y	log(1/IC50)	L_R	B_{5X}	I_{HAL}
1	COCH3	OCH3	CF2H	9.28	4.06	3.07	0
2	COC2H5	OCH3	CF2H	9.23	4.87	3.07	0
3	CON(CH3)2	OCH3	CF2H	9.37	4.77	3.07	0
4	COOCH3	OCH3	CF2H	9.36	4.73	3.07	0
5	COCH3	F	CF2H	9.46	4.06	1.35	1
6[a]	COC2H5	F	CF2H	9.07	4.87	1.35	1
7	CON(CH3)2	F	CF2H	9.46	4.77	1.35	1
8	COOCH3	F	CF2H	9.37	4.73	1.35	1
9[a]	COCH3	Cl	CF2H	8.71	4.06	1.8	1
10	COC2H5	Cl	CF2H	9.37	4.87	1.8	1
11[a]	CON(CH3)2	Cl	CF2H	9.24	4.77	1.8	1
12	COOCH3	Cl	CF2H	9.54	4.73	1.8	1
13	COCH3	N3	CF2H	9.24	4.06	4.18	0
14	COC2H5	N3	CF2H	9.43	4.87	4.18	0
15	CON(CH3)2	N3	CF2H	9.4	4.77	4.18	0
16	COOCH3	N3	CF2H	9.44	4.73	4.18	0
17	COCH3	OCH3	CF3	9.16	4.06	3.07	0
18	COC2H5	OCH3	CF3	9.27	4.87	3.07	0
19	CON(CH3)2	OCH3	CF3	9.17	4.77	3.07	0
20	COOCH3	OCH3	CF3	9.28	4.73	3.07	0
21	COCH3	F	CF3	8.95	4.06	1.35	0
22	COC2H5	F	CF3	8.94	4.87	1.35	0
23	CON(CH3)2	F	CF3	9.12	4.77	1.35	0
24	COOCH3	F	CF3	9.07	4.73	1.35	0
25	COCH3	Cl	CF3	9.07	4.06	1.8	0
26	COC2H5	Cl	CF3	8.95	4.87	1.8	0
27[a]	CON(CH3)2	Cl	CF3	9.35	4.77	1.8	0
28	COOCH3	Cl	CF3	9.17	4.73	1.8	0
29	COCH3	N3	CF3	9.3	4.06	4.18	0
30	COC2H5	N3	CF3	9.4	4.87	4.18	0
31	CON(CH3)2	N3	CF3	9.3	4.77	4.18	0
32	COOCH3	N3	CF3	9.4	4.73	4.18	0
33	H	N3	CF3	9.15	2.06	4.18	0

IC50 označava molarnu koncentraciju kemijskog spoja kojim se inhibira 50 % rasta stanica raka. Za potrebe QSAR modeliranja uzimaju se logaritmirane vrijednosti IC50.. Inače, Corwin Hansch poznat je i kao utemeljitelj QSAR metode, i dugi niz godina primjenjivao ju je u modeliranju biološke aktivnosti molekula ponajviše u medicinskoj kemiji i dizajniranju novih

lijekova. Karakteristično za radove C. Hanscha i suradnika je da su u svim QSAR analizama koristili uglavnom mali broj fizikalno-kemijskih i indikatorskih deskriptora. To je potpuno oprečno od današnjih QSAR studija koje modeliranje vrše na velikom početnom skupu deskriptora. U takvom današnjem pristupu QSAR modeliranju, prethodna eliminacija deskriptora niske složenosti nužan je korak u početnoj fazi modeliranja. Binarni indikatorski deskriptori u prethodnim modelima često imaju veliki nesrazmjer broja molekula u jednoj i drugoj klasi. To je glavni razlog zbog kojeg su izabrani ti skupovi za ilustraciju primjene parametara složenosti varijabli razvijenih u disertaciji. Glavni predmet istraživanja ovog doktorskog rada je postupak izbora deskriptora na način da se zadrže oni prihvatljive složenosti, eliminiraju oni niske složenosti.Za spomenuti skup 33 analoga taksana izračunati su deskriptori: duljina supstituenta R (L_R), širina supstituenta (B_{5X}), dok je I_{HAL} indikatorski deskriptor koji ima vrijednost 1 samo kad je na položaju X supstituent atom halogenog (HAL) elementa flora (F) ili klora (Cl) i da je istovremeno na položaju Y skupina CF2H. U svim ostalim slučajevima indikatorski deskriptor I_{HAL} ima vrijednost 0.

Složenost deskriptora i varijable koja predstavlja biološku aktivnost iz *Tablice 3.12* analizirana je razvijenim mrežnim poslužiteljem *Classification variable complexity parameter estimator* [76], a rezultati analize dani su u *Tablici 3.13*.

Tablica 3.13 Rezultati analize složenosti biološke aktivnosti i deskriptora za skup taksana iz *Tablice 3.12* [67] mrežni poslužiteljem razvijenim u disertaciji [76]

	biološka aktivnost i deskriptori			
	log(1/IC50)	L_R	B_{5X}	I_{HAL}
x	0.5455	0.7273	0.5152	0.2424
X	18	24	17	8
$Q_{2,min}$	0.0909	0.4545	0.0303	0.5152
$Q_{2,max}$	1	1	1	1
$Q_{2,rnd}$	0.5041	0.6033	0.5005	0.6327
$\Delta(Q_2)$	0.9091	0.5455	0.9697	0.4848
$\Delta Q_{2,min}$	-0.4132	-0.1488	-0.4702	-0.1175
$\Delta Q_{2,max}$	0.4959	0.3967	0.4995	0.3673
$\Delta Q_{2,max,norm}$ (%)	99.2 %	79.3	99.9	73.5
$\Delta Q_2 CmpLevel$	≥ 10 %	≥ 10 %	≥ 10 %	≥ 10 %
$logW$	9.01	7.59	9.07	7.14
$logW_{norm}$ (%)	99.0	83.7	100	78.8
$logWCmpLevel$	≥ 10 %	≥ 10 %	≥ 10 %	≥ 10 %

Dobivene razine složenosti (iskazanu u %) dovoljno su visoke za sve parametre, tj. sve su razine iznad 10 % prema oba parametra (i po normiranim parametrima $logW$ i $\Delta Q_{2,max}$). Složenost indikatorskog deskriptora najniža je, iako vrlo značajna te iznosi 79 % prema normaliziranoj entropiji ($logW_{norm}$) i 74 % prema normaliziranom stvarnom doprinosu modela ($\Delta Q_{2,max,norm}$). U slučaju smanjenog skupa, nakon izbacivanja četiri spoja koji su najviše odstupali pa su izbačeni iz modeliranja u [67], dovoljno je promotriti samo promjenu složenosti indikatorskog deskriptora, dok kod ostalih varijabli ne očekujemo značajnije promjene. Kako su od četiri izbačene molekule njih tri s vrijednošću jednakom 1, to je u indikatorskom deskriptoru I_{HAL} preostalo pet molekula s vrijednošću 1, i 24 s vrijednostima jednakim 0. Složenost te varijable je 65 % prema $logW_{norm}$ i

57 % prema $\Delta Q_{2,max,norm}$. To je značajno smanjenje složenosti, iako se promjena omjera klasa 1 i 0 čini malom, te se treba voditi računa o tome kad se radi modifikacija skupa deskriptora na taj način.

Dobivene razine složenosti (iskazanu u %) dovoljno su visoke za sve parametre, tj. sve su razine iznad 10 % prema oba parametra (i po normiranim parametrima $logW$ i $\Delta Q_{2,max}$). Složenost indikatorskog deskriptora najniža je, iako vrlo značajna te iznosi 79 % prema normaliziranoj entropiji ($logW_{norm}$) i 74 % prema normaliziranom stvarnom doprinosu modela ($\Delta Q_{2,max,norm}$).

Četvrti skup deskriptora za analoga 22 taksana , a priređen je u radu Verme i Hanscha [67] (Tablica 1 iz toga rada). Skup sadrži tri deskriptora od kojih je jedan indikatorski deskriptor koji ima vrijednost 1 samo ako je supstituent X cikloalkilna skupina, a preostala dva su hidrofobnost (π_X) i molarna refraktivnost (MR_X) supstituenta X (*Tablica 3.14*). Na tom skupu spojeva modelirala se njihova citotoksičnost log(1/IC50) prema staničnim linijama oznake HCT-116 iskazana u jedinici 10^{-9} mola (nM), a QSAR model dan je jednadžbom (1) u tom radu [67].

Tablica 3.14 Skup biološke aktivnosti i deskriptora za 22 spoja iz skupine pacitaksela (*pacitakseli22*) u QSAR modeliranju iz rada Verme i Hanscha [67]

No.	X	log(1/IC50)	π_X	MR_X	I_{CYALK}
1	CH3	8.62	0.7	0.46	0
2	C6H5	6.39	1.91	2.51	0
3	4-F-C6H4	6.1	2.05	2.53	0
4	CH2F	8.15	0.37	0.48	0
5	CCl3	8.4	2.58	1.94	0
6	C2H5	8.7	1.23	0.93	0
7	CH]CH2	8.22	0.85	0.98	0
8	(CH2)2CH3	8.96	1.76	1.39	0
9	CH(CH3)2	8.3	1.54	1.39	0
10	C(CH3)=CH2	8.35	1.16	1.44	0
11	trans-CH=CHCH3	8.64	1.38	1.44	0
12	Cy-C3H5	9	1.28	1.25	1
13	(CH2)3CH3	8.7	2.28	1.86	0
14	Cy-C4H7	8.82	1.61	1.68	1
15	(CH2)4CH3	8.22	2.81	2.32	0
16	Cy-C5H9	8.7	2.17	2.14	1
17	OCH3	8.7	0.68	0.62	0
18	OCH2CH3	9	1.21	1.08	0
19	O(CH2)2CH3	8.59	1.74	1.54	0
20	NH-Cy-C4H7	8.04	1.55	2.05	1
21	Imidazole	6.1	0.18	1.73	0
22	Aziridine	7.81	0.87	1.12	0

Složenost deskriptora i varijable koja predstavlja biološku aktivnost iz *Tablice 3.14* analizirana je mrežnim poslužiteljem razvijenim u disertaciji [76], a rezultati analize dani su u *Tablici 3.15*.

Tablica 3.15 Rezultati analize složenosti biološke aktivnosti i deskriptora/varijabli skupa *pacitakseli* iz *Tablice 3.14* [69] mrežnim poslužiteljem razvijenim u disertaciji [76]

	log(1/IC50)	π_X	MR_X	I_{CYALK}
x	0.73	0.5	0.45	0.18
X	16	11	10	4
$Q_{2,min}$	0.45	0	0.09	0.64
$Q_{2,max}$	1	1	1	1
$Q_{2,rnd}$	0.60	0.50	0.50	0.70
$\Delta(Q_2)$	0.55	1	0.91	0.36
$\Delta Q_{2,min}$	-0.15	-0.5	-0.41	-0.07
$\Delta Q_{2,max}$	0.40	0.5	0.50	0.30
$\Delta Q_{2,max,norm}$ (%)	79	100	99	60
$\Delta Q_2 CmpLevel$	≥ 10 %	≥ 10 %	≥ 10 %	≥ 10 %
$logW$	4.9	5.8	5.8	3.9
$logW_{norm}$ (%)	83	100	99	66
$logWCmpLevel$	≥ 10 %	≥ 10 %	≥ 10 %	≥ 10 %

Dobivene razine složenosti dovoljno su visoke za sve parametre, tj. sve su razine iznad 10 % prema oba normalizirana parametra – i $logW_{norm}$ i $\Delta Q_{2,max,norm}$. Složenost indikatorskog deskriptora najniža je, iako vrlo značajna, i iznosi. 66 % prema normaliziranoj entropiji ($logW_{norm}$) i 60 % prema normaliziranom stvarnom doprinosu modela ($\Delta Q_{2,max,norm}$).

Za usporedbe normiranih vrijednosti $logW$ i $\Delta Q_{2,max}$ parametara, napravljena je simulacija kojom su se računali uvjeti dostizanja pragova složenosti (udio klase 1 (x) u varijabli) na varijablama koje imaju od 15 do 2000 vrijednosti/elemenata (tj. $N \in 15, 2000$), a rezultati su dani u *Prilogu E_3.7*. Usporedba u tom rasponu N pokazuje da normirani parametar $logW_{norm}$ postiže razine složenosti 1, 2.5, 5, 7.5 i 10 % uvijek ranije nego $\Delta Q_{2,max,norm}$ ili, u graničnim slučajevima, za isti x (udio klase 1). Vrijednosti x za koje se u nekim slučajevima podjednako brzo dostižu iste razine značajnosti po obje mjere složenosti, nalaze se u intervalu $N \leq 398$. Takvih je slučajeva ukupno 106, postiže se samo za razine složenosti 1 % i, rjeđe, za razine složenosti 2.5 %.

3.6.3 Primjena u analizi deskriptora/varijabli izračunanih na proteinskim sekvencama

Pored primjene na skupovima organskih kemijskih spojeva (ili kako se katkad nazivaju – malim molekulama), pokazat će se primjena rezultata iz disertacije u analizi složenosti varijabli izračunanih na proteinskim sljedovima (sekvencama), kao većim kemijskim/biokemijskim molekularnim strukturama. Razvijena je aplikacija ProtSeqAnalizer [82] (*Prilozi 3.45* i *E_3.3*) za izračun deskriptora iz proteinskih sljedova, a njena primjena ilustrirat će se u analizi baze DADP [65] s 568 antimikrobnih peptida.

Aplikacija ProtSeqAnalizer [82] (*Prilozi 3.45* i *E_3.3*) računa kao deskriptore frekvencije pojavljivanja u proteinima:

(1) motiva – skupina susjednih aminokiselina poput GXXG (*skup*_GXXG, 45 deskriptora/motiva) ili GXXXG (*skup*_GXXXG, 76 deskriptora/motiva), gdje je 'G' oznaka za aminokiselinu glicin, dok je 'X' oznaka za bilo koju aminokiselinu;

(2) pojedinih aminokiselina (*skup*_aminokiseline, 20 deskriptora)

(3) parova susjednih aminokiselina ('dipeptida') u proteinima (*skup*_parovi-1, 370 deskriptora) uzimajući u obzir poredak (AK nije isto kao i KA)

(4) parova susjednih aminokiselina ('dipeptida') u proteinima (*skup*_parovi-2, 203 deskriptora) ne uzimajući u obzir poredak (AK i KA zajedno se zbrajaju i čine jedan deskriptor)

Nadalje, programima razvijenim za računanje deskriptora (malih) organskih molekula računaju se još dva skupa:

(5) deskriptori temeljeni na konceptu povezanosti atoma u molekulama – modificirani zagrebački indeksi/deskriptori (*skup*_ zagrebački, 22 deskriptora - topološki deskriptori temeljeni na reprezentaciji molekule matematičkim grafom) računani aplikacijom razvijenom u disertaciji (*Prilog E_3.2* i [53])

(6) deskriptori izračunani programom Dragon 3.5 (skup_Dragon, 1205 deskriptora) [6]. Skup sadrži razne skupine molekularnih deskriptora, od konstitucijskih, preko topoloških, informacijskih, itd. Taj je program ponajčešće korišten u današnjim QSAR modeliranjima.

Ukupno 1941 deskriptor svrstan je u šest skupova deskriptora. Prije analize svi su deskriptori dihotomizirani s obzirom na srednju vrijednost i vrijednosti svakog deskriptora samo su 1 i 0. Rezultati analize njihove složenosti mrežnim poslužiteljem razvijenim u disertaciji [76] zbirno su prikazani u *Tablici 3.16.*

Tablica 3.16 Zbirni rezultati obrade složenosti varijabli/deskriptora skupa peptida baze DADP [65] mrežnim poslužiteljem [76] razvijenim u disertaciji

razina složenosti	broj proteina ispod razine složenosti		kumulativni zbroj proteina po razinama složenosti	
	$logW_{norm}$	$\Delta Q_{2,max,norm}$	$logW_{norm}$	$\Delta Q_{2,max,norm}$
< 1 %	0	97	0	97
1 – 2.5 %	97	92	97	189
2.5 –5 %	92	104	189	293
5 – 7.5 %	52	58	241	351
7.5 – 10 %	52	43	293	394
≥ 10 %	1648	1547	1941	1941

Broj deskriptora kojima je razina složenosti < 1 % prema normaliziranoj entropiji $logW_{norm}$ bitno je manju nego kad se razina složenosti iskazuje preko maksimalnog normaliziranog doprinosa modela iznad nasumične točnosti ($\Delta Q_{2,max,norm}$). To potvrđuje ranije opisanu zakonitost prema kojoj parametar $logW$ za sve udjele klase 1 raste sporije nego parametar $\Delta Q_{2,max}$. Razlike između ovih dvaju mjera izraženih u postotcima najveće su kod malih udjela klase 1 (x) i složenostima manjim od 1 %, a približno se izjednačuju polazeći već od složenosti > 1 %. Ukoliko bi se ova metoda iskoristila za odabir varijabli tako da se npr. odabere razina složenosti od ≥ 10 %, tada bi od 1941 deskriptora njih 394 bilo odbačeno prema $logW_{norm}$, i 293 prema parametru složenosti $\Delta Q_{2,max,norm}$.

GXXG motiv čest je kod antimikrobnih peptida izoliranih iz žabe. Njegova učestalost pojavljivanja u skupu DADP polipeptida [65] obrađena je i spremljena u datoteke koje imaju u nazivu 'GXXG' u *Prilogu E_3.8*). Obrada sekvenci koje sadržavaju taj motiv prikazana je u *Tablici 3.17*.

Tablica 3.17 Značajnost GXXG deskriptora za 568 peptida iz baze DADP [65] prema parametrima $\Delta Q_{2,max,norm}$ i $logW_{norm}$ izračunana mrežnim poslužiteljem razvijenim u disertaciji [76]

x (%)	X	deskriptor (motiv)	$\Delta Q_{2,max,norm}$ (%)	$\Delta Q_{2,max,norm}$ razina složen.	$logW_{norm}$ (%)	$logW_{norm}$ razina složenosti
4.2	24	GLSG	16.2	≥ 10 %	24.8	≥ 10 %
2.3	13	GKTG	9.0	≥ 7.5 %	15.3	≥ 10 %
2.3	13	GILG	9.0	≥ 7.5 %	15.3	≥ 10 %
1.8	10	GIGG	6.9	≥ 5 %	12.4	≥ 10 %
1.8	10	GVLG	6.9	≥ 5 %	12.4	≥ 10 %
1.4	8	GLLG	5.6	≥ 5 %	10.3	≥ 10 %
1.1	6	GLFG	4.2	≥ 2.5 %	8.1	≥ 7.5 %
1.1	6	GKAG	4.2	≥ 2.5 %	8.1	≥ 7.5 %
1.1	6	GKVG	4.2	≥ 2.5 %	8.1	≥ 7.5 %
0.9	5	GKFG	3.5	≥ 2.5 %	6.9	≥ 5 %
0.7	4	GLVG	2.8	≥ 2.5 %	5.7	≥ 5 %
0.7	4	GFLG	2.8	≥ 2.5 %	5.7	≥ 5 %
0.5	3	GNTG	2.1	≥ 1 %	4.4	≥ 2.5 %
0.5	3	GLTG	2.1	≥ 1 %	4.4	≥ 2.5 %
0.4	2	GMLG, GAFG [a]	1.4	≥ 1 %	3.1	≥ 2.5 %
0.2	1	GGKG, GNMG [b]	0.7	< 1 %	1.6	≥ 1 %

[a] Identične sve numeričke vrijednosti složenosti imaju i ovi motivi: GLKG, GIFG, GVSG, GGGG;
[b] Identične sve numeričke vrijednosti složenosti ima i ovih 23 motiva: GPHG, GING, GVAG, GALG, GAAG, GIVG, GIHG, GFKG, GVKG, GLRG, GLAG, GKMG, GIKG, GLGG, GVIG, GGRG, GGSG, GRGG, GRRG, GRSG, GSGG, GSRG, GRHG

Aminokiselinski motivi oblika GXXG poput ovih u *Tablici 3.17* analizirani su kao bitni u modeliranju antimikrobne aktivnosti peptida [65]. Samo jedan deskriptor/motiv ima složenost veću od 10 %, a njih 6 veću od 5 % prema $\Delta Q_{2,max,norm}$, dok šest motiva ima složenost veću od 10 % i njih 12 veću od 5 % prema normaliziranoj entropiji $logW_{norm}$. Ukoliko se uzme 1 % kao minimalni prag prihvatljive složenosti, onda prema $\Delta Q_{2,max,norm}$ 25 motiva/deskirptora treba biti odbačeno od njih 45, dok prema $logW_{norm}$ svih tih 25 motiva/deskriptora ima složenost između 1 i 2.5 %. Tih 25 motiva ima samo jednu vrijednost jednaku 1 i 567 vrijednosti koje su jednake 0. U slučaju kada i ta jedna vrijednost ne bi bila jednaka 1, taj bi deskriptor imao identičnih svih 568 vrijednosti (koje su sve 0) i bio bi zacijelo odbačen kao potpuno monoton (neinformativan).

Maksimalno mogući broj motiva oblika GXXG je 400 (na svakom mjestu X može se izmijeniti do 20 aminokiselina), međutim samo 45 peptida ima barem jedan takav motiv. Razine složenosti za $N = 568$ peptida u ovisnosti o minimalnom broju proteina (X) koji imaju motiv GXXG (broj peptida klase 1 (X)) dane su u *Tablici 3.18.*

Iz *Tablice 3.18* vidi se da razina (prag) složenosti prema $logW_{norm}$ u ovisnosti o (minimalnom) broju peptida klase 1 koji imaju barem jedan motiv GXXG raste brže nego složenost prema $\Delta Q_{2,max,norm}$, što ukazuje da je entropija manje strog kriterij složenosti.

Tablica 3.18 Razine složenosti varijabli u ovisnosti o normaliziranim parametrima složenosti i o minimalnom broju elemenata (peptida) klase 1 za varijablu s N = 568 vrijednosti

minimalan broj peptida klase 1 (X) dovoljan za postizanje razina složenosti		razina složenosti
$\Delta Q_{2,max,norm}$ (%)	$logW_{norm}$ (%)	
1	0	< 1 %
2	1	≥ 1 %
4	2	≥ 2.5 %
8	4	≥ 5 %
11	6	≥ 7.5 %
15	8	≥ 10 %

Međutim, logično je uzeti da deskriptor mora imati minimalno 5 % složenosti po oba normirana parametra složenosti, što bi značilo da deskriptor mora imati odabrani strukturni motiv GXXG u minimalno 8 od 567 peptida (a u preostalih 559 peptida vrijednost 0). Prema tom kriteriju, u slučaju deskriptora iz skupine GXXG za 568 peptida iz modeliranja bi se uklonilo 40 od 45 deskriptora.

U nastavku bit će ukratko opisani rezultati analize 370 deskriptora dipeptida , koji opisuju poredak aminokiselina (*Prilog E_3.8*), datoteka aa2.csv.col_formula.csv. Prema normaliziranoj entropiji $logW_{norm}$ 233 dipeptida imaju složenost veću od 10 %, 65 je u rasponu 5 % do 10 %, 72 ima složenost manju od 5 %, dok prema $\Delta Q_{2,max,norm}$ odgovarajući brojevi dipeptida manji su i iznose 170 (≥10 %), 63 (5-10 %) i 137 (< 5 %).

Analiza 1205 Dragon deskriptora [6] i 22 deskriptora iz skupine modificiranih zagrebačkih indeksa [6] (*Prilog E_3.8*), pokazala je kako deskriptori imaju složenost ≥10% prema oba normalizirana parametra. Pošto su molekule proteina velike, niti jedan deskriptor se ne mora isključiti. U takvim strukturama, svaki od strukturnih detalja (opisanih deskriptorima) ima frekvenciju pojavljivanja u najmanje 15 peptida iz skupa od 568 peptida, tj. ima vrijednosti koje pripadaju klasi 1 (nakon dihotomizacije deskriptora s obzirom na srednju vrijednost). Sve varijable iz ova dva skupa imaju razinu složenosti veću od 10 % (prema *Tablici 3.18*).

3.6.4 Primjena izvedenih parametara u procjeni kvalitete i rangiranju modela

Parametri Q_2, ΔQ_2, MCC i $F1$ upotrebljeni za analizu složenosti u disertaciji za analizu složenosti klasifikacijskih varijabli s dva stanja, uporabljeni su za analizu prediktivne kvalitete modela (metoda) s natjecanja u predviđanju genskih mutacija koje dovode do tumora. Parametri MCC i $F1$ često se rabe u analizi kvalitete modela za predviđanje na nepoznatom skupu. Neki autori daju prednosti koeficijentu korelacije MCC [37,78], iako je vrlo jednostavno uočiti kako parametar nije prikladan u procjeni ročnosti predviđanja modela jer je neosjetljiv na konstantni pomak vrijednosti varijabli između kojih se računa korelacija [35]. Bez obzira koliki bio taj konstantni pomak (što je, stvarno, pogreška modela), vrijednost koeficijenta korelacije uvijek je ista. Nadalje, $F1$ nije prikladan ukoliko broj točnih predviđanja većinske klase 0 nije beskonačan (tj. ukoliko nije $n + u \to \infty$, odnosno $n \to \infty$).

Korisna svojstva parametra ΔQ_2 bit će ilustrirana u primjeni rangiranja 70 modela na skupu podataka priređenom u radu Lučić i dr. [35] i Cooper i dr. [79] (Tablica S8 u dodatnoj datoteci broj

9). Tih 70 modela odnose se na modele iz završne faze prediktivnog natjecanja (IS3). Skup podataka koji je modeliran u tom natjecanju sadrži 24687 slučajeva (gena), među kojima je 7903 (32 %) pozitivne (somatska mutacija gena povezana s nastankom tumora), a preostalih 16784 (68 %) negativne klase (somatska mutacija gena koja ne izaziva nastanak tumora) [80]. Parametar $F1$ korišten je kao glavni kriterij ocjenjivanja (metoda rangiranja) [79].

Tablica sa 70 modela nalazi se u *Prilozima 3.46 i 3.47*. Rezultati u tablici (*Prilozi 3.46 i 3.47*) poredani su prema padajućim vrijednostima parametra ΔQ_2. U toj tablici 10 najbolje rangiranih modela prema Q_2, MCC i $F1$ nalaze se unutar 15 najboljih modela rangiranih prema ΔQ_2. Nadalje, 20 najbolje rangiranih modela prema svakom od četiri parametra nalazi se među 21 najboljim modelom rangiranim prema ΔQ_2 (stvarni doprinos modela iznad nasumične točnosti/korelacije). Srednja apsolutna razlika rangova prema Q_2, MCC i $F1$ u odnosu na rangove prema ΔQ_2 u najboljih 15 modela iznose 5.9, 6.3 i 4.3, a za kompletnu listu od 70 modela iznose 4.1, 4.0 i 2.9.

Najbolja tri modela prema ΔQ_2 (X2463247, X2478107, i X2453885) rangirani su kao 1, 2 i 3, dok su ti isti modeli prema vrijednostima parametra $F1$ rangirani (redno) kao 5., 10. i 6., te na 10., 5. i 15. mjesto prema parametru MCC.

Odnosi n/p slični su kod svih najboljih modela, ali najmanje vrijednosti odnosa pogrešaka o/u su kod modela koji su rangirani kao najbolji prema ΔQ_2 - i taj je omjer u rasponu od 2 do 3.2. To ukazuje da parametar favorizira modele s ravnotežnim pogreškama, tj. modele koji su bliže uravnoteženim modelima. Takva vrsta modela ponajbolja je vrsta modela, jer ujednačuje i uravnotežuje vjerojatnost pogrešaka predviđanja obje klase. Odgovarajući raspon za prva tri modela prema $F1$ je od 4.4 do 9.8 i sličan je rasponu vrijednosti o/u u modelima rangiranima kao najbolji prema Q_2 i MCC. Stoga se može reći kako $F1$ i MCC [73,81] na ovom primjeru favoriziraju predviđanje jedne od dviju klasa, dok ΔQ_2.ukazuje da su bolji oni modeli koji ujednačuju pogreške u i o. Nadalje, omjer o/u kod najboljih modela prema ΔQ_2 blizak je omjeru n/p. Takvi modeli povećavaju stvarni doprinos modela iznad nasumične točnosti, i stoga se –razumljivo – mogu smatrati boljima.

Parametar $F1$ nije odgovarajuća mjera za klasu 1 kad je ona većinska, tj. tad bi se trebale koristiti dvije odvojene varijante ovog parametra za dvije klase. Vrijednosti parametara $F1$ i MCC pokazuju jaku osjetljivost o raspodjeli podataka, tj. o omjeru udjela klasa. Osim toga, za razliku od parametara $F1$ i MCC, parametar ΔQ_2 definiran je za svaki skup vrijednosti iz tablice pogrešaka, pa prema tome i za svaki mogući model M i eksperimentalnu varijablu E, i pokazuje linearno proporcionalnu vrijednost u odnosu na promjene vrijednosti elemenata matrice pogrešaka.

3.7 Poopćenje rezultata dobivenih za izmjenjive varijable

Izrazi minimalnih i maksimalnih karakterističnih vrijednosti raznih parametara izvedeni za izmjenjive varijable u ovisnosti o udjelu klase 1 (x), analogno bi trebali vrijediti i za općenite varijable. U disertaciji se pokazalo da je tako nešto moguće s pomoću supstitucijskih tablica koje ovise o dva dijela klase – jedan za eksperimentalni udio klase 1 (x), a drugi za udio klase 1 predviđen modelom (y). Takvi izrazi omogućit će bolju i vjerodostojniju provjeru kvalitete modela kad se kvaliteta modela iskazuje proizvoljnim parametrom (a postoji jako veliki broj parametara – mjera kvalitete – koji se koriste u istraživanjima za procjenu kvalitete modela [33,78]).

Nadalje, analogno su uvedene supstitucijske tablice (i simulacijski provjerena njihova ispravnost) za računanje prosječnih nasumičnih vrijednosti parametara razmatranih u disertaciji.

Spomenute supstitucijske tablice izvedene su u disertaciji, iako se prvobitno nije očekivalo da bi to bilo moguće. Stoga, to predstavlja dodatni originalni doprinos najavljenim i prvobitno planiranim rezultatima u disertaciji.

3.7.1 Minimalne i maksimalne vrijednosti parametara kvalitete općenitih binarnih varijabli

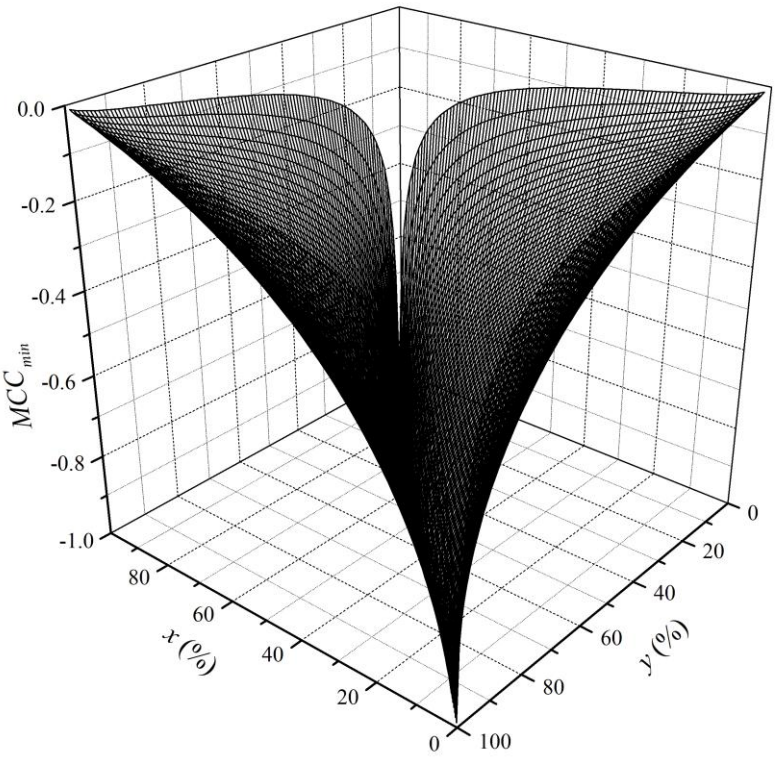
U formulama supstitucijske tablice (*Prilog 3.36*) uvedene su dvije varijable udjele klase 1, pa je tako udio x kod neizmjenjivih varijabli oznaka za udio klase 1 u eksperimentalnoj varijabli E, dok se nova oznaka y uvodi za udio klase 1 u varijabli M.

Udjeli x i y kod neizmjenjivih varijabli računaju se pomoću formula (3.59).

$$x = \frac{p+u}{N}, \qquad y = \frac{p+o}{N}, \forall x, y \in [0,1] \tag{3.59}$$

Uvrštavanjem elemenata matrice pogrešaka p, n, u i o iz supstitucijske tablice varijabli (izmjenjivih i svih ostalih) iz *Priloga 3.36* u formule parametara Q_2, $Q_{2,rnd}$, ΔQ_2, MAE, s, MCC, $F1$ ili κ postaju funkcije dviju varijabli udjela klasa, a njihove su vrijednosti prikazane u *Prilogu 3.37*.

Prikaz minimalne vrijednosti koeficijenta korelacije (MCC) u općenitom slučaju binarnih varijabli kad ona ovisi o udjelima klase 1 u obje varijable (E i M), dan je na *Slici 3.31*.



Slika 3.31 Ovisnost minimuma parametra MCC za općeniti slučaj dviju varijabli E i M u ovisnosti o udjelima klase 1 u njima (redno: x i y)

Izračuni potrebni za izradu ove slike izrađeni su u jeziku R [58], dok je sama slika izrađena programom Origin. [60]

3.7.2 Prosječne nasumične vrijednosti parametara kvalitete općenitih binarnih varijabli

Prosječne nasumične vrijednosti parametara vrijednosti su koje se mogu dobiti supstitucijama za elemente matrice pogrešaka danih u *Tablici 3.19* u originalne izraze za izračun parametara (mjera) kvalitete modela.

Tablica 3.19 Tablica supstitucija za računanje prosječnih nasumičnih vrijednosti parametara

Parametar	Supstitucija
p	$\frac{(p+o)(p+u)}{N}$
n	$\frac{(n+o)(n+u)}{N}$
o	$\frac{(n+o)(p+o)}{N}$
u	$\frac{(n+u)(p+u)}{N}$

***** N – ukupni broj podataka u varijabli, p – broj točno predviđenih podataka klase 1, n – broj točno predviđenih podataka klase 0, o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1

Do sada je u literaturi bio poznat samo izraz za izračun $Q_{2,rnd}$, tj. prosječne nasumične vrijednosti parametra točnosti Q_2 [34,35]. Izvodi za većinu prosječnih nasumičnih vrijednosti ostalih parametara kvalitete nisu bili objavljeni u literaturi. Također, izvodi prosječnih nasumičnih vrijednosti općenitih parametara danih jednadžbama (2.5) do (2.9) (tj. za varijable koje nisu izmjenjive i kod kojih nije $u = o$) mogu se dobiti uporabom supstitucijskih vrijednosti za p, n, u i o iz *Tablice 3.19.* Njihove prosječne nasumične vrijednosti dane su u *Tablici 3.20.*

U ranijim radovima prosječna nasumična točnost modela nazivana je najvjerojatnijom nasumičnom točnosti modela, ali se pokazalo da ta vrijednost odgovara srednjoj vrijednosti [34,35]. U slučaju parnog broja podataka u varijabli, pokazalo se da ta vrijednost ne postoji kao stvarna vrijednost parametra, nego samo kao srednja vrijednost parametra, pa je naziv iz ranijih radova [34,35] promijenjen (korigiran), što predstavlja vrijedan rezultat dobiven u disertaciji.

Tablica 3.20 Prosječne nasumične vrijednosti parametara kvalitete

parametar	prosječna nasumična vrijednost parametra
Q_2	$Q_{2,rnd} = \frac{(p+u)(p+o)+(n+u)(n+o)}{N^2}$
ΔQ_2	0
$Q_{2,rnd}$	-
MAE	$\frac{(p+o)(n+o)+(p+u)(n+u)}{N^2}$
s	$(\frac{1}{N})\ \overline{(p+o)(n+o)+(p+u)(n+u)}$
MCC	0
$F1$	$\frac{2\ o+p\quad p+u}{o+2p+u\ \ N}$
κ	0

Svaki parametar u *Tablici 3.20* predstavlja stvarnu srednju vrijednost parametra, a svaki od tih rezultata provjeren je simulacijama (za 100.000 simulacija parova varijabli).

3.7.3 Izvod standardne devijacije i standardne pogreške srednje vrijednosti parametara kvalitete binarnih varijabli

Od dodatnih izvoda parametara, važno je samo napomenuti standardnu devijaciju (σ) i standardnu pogrešku srednje vrijednosti (*SE*) klasifikacijskih varijabli s dva stanja . Oba parametra moguće je dobiti iz udjela klase 1, bez da se podaci u cijelosti koriste kao argument funkcija, tj. izvodi dokazuju da je standardnu devijaciju moguće dobiti iz udjela klase 1 (x), i isto tako iz elemenata tablice pogrešaka $p, n, u,$ i o [83].

Izvodi parametara σ i SE nalaze se u prilozima (*Prilozi 3.38 – 3.40*), a njihove su formule (3.60) do (3.62). Oznaka E u indeksu devijacije označava da je riječ o eksperimentalnoj varijabli.

$$\sigma = \overline{xN\frac{(1-x)}{N-1}} \forall x \in [0,1] \tag{3.60}$$

$$\sigma_E = \overline{\frac{(p+u)(N-p-u)}{N(N-1)}} = \overline{\frac{(p+u)(n+o)}{N(N-1)}} \tag{3.61}$$

$$SE = \frac{\sigma}{\overline{N}} = \overline{x\frac{(1-x)}{(N-1)}} \tag{3.62}$$

$$SE_E = \frac{1}{N}\ \overline{\frac{(p+u)(N-p-u)}{N-1}} = \frac{1}{N}\ \overline{\frac{(p+u)(n+o)}{N-1}} \tag{3.63}$$

U slučaju da se radi o izmjenjivim varijablama σ (i SE) će dati jednak rezultat na objema varijablama, a u slučaju generalnih varijabli, potrebno je u jednadžbi parametar u zamijeniti s parametrom o (formula (3.61)).

4. RASPRAVA

U literaturi postoje brojni algoritmi za procjenu složenosti strukture molekula koji se temelje na složenijim konceptima poput: (1) složenosti grafa koji predstavlja strukturu molekule; (2) povezanosti atoma u molekuli; (3) stupnju grananja strukture s obzirom na valenciju atoma (čvorova grafa), itd. [24] U svim takvim analizama postoji puno proizvoljnih funkcionalnih ovisnosti između raznih strukturnih elemenata i njihovih pojednostavljenih reprezentacija (poput atom – čvor grafa; kemijska veza – brid grafa, itd.). Najčešće su funkcionalne ovisnosti oblika Shannonove entropije $n \cdot \ln(n)$ [69,70], gdje n predstavlja broj veza koji neki čvor (atom) može formirati, broj atoma, broj prstena u molekuli, itd. To su nešto drugačiji i kompliciraniji vidovi razmatranja složenosti u odnosu na istraživanja provedena u disertaciji, a tiču se QSAR modeliranja u kemiji, i odnose se, u pravilu, na složenost molekularne strukture [24]. Nadalje, računa se i entropija i degeneracija (visoka degeneracija znači nisku entropiju) skupa deskriptora ili degeneracija samih pojedinačnih deskriptora [84]. Taj pristup može se izdvojiti kao onaj koji je najbliži istraživanjima u disertaciji, iako nije fokusiran na analizu pojedinačnog klasifikacijskog deskriptora s dva stanja (binarni deskriptor). Nadalje, u literaturi [84] postupci nisu planirani niti provedeni s ciljem definiranja jasnih preporuka i dobivanja rezultata u vezi složenosti/entropije pojedinačnog deskriptora/varijable koji bi se mogli koristiti za određivanje kriterija za uključivanje ili isključivanje pojedinog deskriptora iz QSAR modeliranja kao slabo informativnog, ili iz konačnih QSAR modela. Prema dostupnoj literaturi, istraživanja provedena u disertaciji i dobiveni rezultati originalni su doprinos analizi složenosti varijabli.

4.1 Karakteristične vrijednosti parametara kvalitete modela

4.1.1 Stvarna točnost uravnoteženih modela

Do temeljne zamisli za provođenje ovog istraživanja došlo se analizom parametra točnosti Q_2, nasumične točnosti $Q_{2,rnd}$ i njihove razlike ΔQ_2 (nazvane „stvarni doprinos modela iznad nasumične točnosti“) u slučaju uravnoteženog modela/predviđanja [34,35], kako je definirano u tom radu. Balansirano predviđanje idealni je oblik predviđanja koje u cijelosti reproducira raspodjelu podataka u varijabli (tj. eksperimentalne aktivnosti) koji se modeliraju. To znači da ako je model razvijen na eksperimentalnoj binarnoj klasifikacijskoj varijabli s N vrijednosti/molekula od kojih njih X ima vrijednost 1 (i udio x) i $N - X$ vrijednost 0 (i udio $1 - x$), uravnoteženi model predviđa podjednak broj vrijednosti/molekula u klasi 1 i u klasi 0. To se postiže postupkom izbora modela koji se ugađa tako da reproducira raspodjelu podataka na kojima se razvija i optimira.

Maksimalni mogući stvarni doprinos uravnoteženog modela iznad nasumične točnosti jednak je $(\Delta Q_2)_{max} = (Q_2)_{max} - Q_{2,rnd}$=$1 - Q_{2,rnd}$. Analiza stvarnog maksimalnog stvarnog doprinosa modela iznad nasumične točnosti (ΔQ_2) pokazala je da taj doprinos ovisi samo o udjelima klase 1 u eksperimentalnoj varijabli i u varijabli predviđenoj modelom (koji su jednaki - i označavaju se s x):

$$(\Delta Q_2)_{max} = 1 - \ 2x^2 - 2x + 1 \ = -2x^2 + 2x, \forall x \in [0,1] \tag{4.1}$$

4.1.2 Složenost uravnoteženog modela i analogija sa složenošću varijable

Kada je x mali, ili kada je blizu 1, najveći mogući (maksimalni) stvarni doprinos modela ΔQ_2 iznad nasumične točnosti postaje jako mali. U graničnom slučaju uravnoteženih modela kada je $x = 0$ (odnosno $x = 1$), sve vrijednosti i u experimentalnoj varijabli (E) i u onoj predviđenoj modelom (M) jednake su 0 (odnosno 1), proizlazi da je $\Delta Q_2 = 0$. U najkompleksnijem slučaju kada i eksperimentalna varijabla E i predviđanje uravnoteženim modelom (modelna varijabla M) sadrže podjednak udio klase 1 ($x = 1/2$), a time i klase 0 ($1 - x = 1/2$), $(\Delta Q_2)_{max} = 1/2$, što je maksimalna moguća vrijednost tog parametra.

Ova dva rubna slučaja upućuju na to da parametar ΔQ_2 korelira s varijabilnošću odnosno složenošću varijabli E i M kod uravnoteženog modela. Kad varijabilnost postane nula, model ne može doprinijeti nikakvu korisnu informaciju iznad nasumične točnosti, jer maksimalna moguća točnost postane 1, jednako kao i (prosječna) nasumična točnost. S druge, kad je varijabilnost maksimalna, i maksimalni mogući stvarni doprinos modela je maksimalan:

$$(\Delta Q_2)_{max} = 1/2 \tag{4.2}$$

Ova razmatranja dala su ideju kako bi maksimalni stvarni doprinos modela mogao biti temelj za analize složenosti pojedinačnih varijabli.

Kod uravnoteženih modela, i eksperimentalna binarna klasifikacijska varijabla E i binarna klasifikacijska varijabla koja predstavlja predviđanje uravnoteženim modelom (modelna varijabla M) imaju identične raspodjele (udjele klasa 1 i 0). Može se zamisliti varijablu E kao općenitu binarnu klasifikacijsku varijablu s vrijednostima 1 i 0 (pripadnost klasi 1 i klasi 0) u stalnom poretku. Predviđanje uravnoteženim modelom analogno je nekoj varijabli M koja ima isti broj vrijednosti koje su jednake 1 i 0 kao i varijabla E, samo su drugačije poredane u odnosu na originalni poredak eksperimentalne varijable E (koja je poslužila za razvoj i ugađanje modela). S obzirom na to da, model ima uvijek neku pogrešku, usporedba vrijednosti varijabli E i M dat će informaciju o ukupnom slaganju tih dviju varijabli. Ukoliko su varijable E i M slabo varijabilne, tj. ukoliko im je udio klase 1 mali (~ 0) ili velik (~ 1.0), bit će jako mali broj mogućih različitih rasporeda varijable M, pa će i razvijeni uravnoteženi model imati manju složenost. Manja složenost uravnoteženog modela znači i manju složenost početne varijable E, i modelne varijable M.

4.1.3 Permutacijske analize i karakteristične vrijednosti parametara kvalitete

S obzirom da imaju iste udjele klasa 1 i 0, modelna varijabla M, kao predviđanje dobiveno uravnoteženi, modelom razvijenim na eksperimentalnoj varijabli E, može se promatrati kao jedna je od mnogo mogućih neidentičnih permutacija varijable E. Skup svih takvih neidentičnih permutacija varijable M definira prostor unutar kojega se nalaze sve moguće optimizacije modela koje se provode u postupku učenja - pri čemu se model optimira i prilagođuje eksperimentalnoj varijabli E. Za svaku tako dobivenu varijablu M može se računati podudarnost njenih vrijednosti u usporedbi s varijablom E, čije su vrijednosti uvijek u stalnom poretku. Tijekom rada na interpretaciji rezultate otkriveno je da taj koncept uveden u disertaciji odgovara matematičkoj teoriji izmjenjivih varijabli [39,40]. Tako definiran koncept izmjenjivih varijabli E i M može se primijeniti na općenitu binarnu klasifikacijsku varijablu E u stalnom poretku i na njene neidentične permutacije M. Intuitivno je

jasno da je analiza podudarnosti između svih mogućih takvih parova varijabli (E, M), te njihov ukupni broj, u vezi sa složenošću varijable E.

Može se zamisliti da se među svim parovima izmjenjivih varijabli E i M (pri čemu je varijabla E uvijek fiksna - u stalnom poretku) nalaze i permutacije varijable M koje imaju najlošiju (minimalnu) i najbolju (maksimalnu) podudarnost s varijablom E. Podudarnost između varijabli E i M nastoji se kvantificirati nekim brojčanim pokazateljem. Ta se podudarnost najprije kvantificira elementima tablice pogrešaka (*engl*. confusion table), koja se sastoji od četiri cjelobrojne vrijednosti p, n, u i o [33,34,78] objašnjene u dijelu 2.1.2 i u Tablici 2.1. Shodno tome, p (odnosno n) predstavljaju slaganje između vrijednosti 1 (odnosno 0) varijabli E i M, tj. broj slučajeva kad je pogreška jednaka 0. Nadalje, u (odnosno o) predstavljaju brojeve slučajeva kada je vrijednost 1 (odnosno 0) u varijabli E predviđena kao 0 (odnosno 1) u varijabli M. Najčešće se podudarnost brojčano iskazuje parametrom točnosti (Q_2) koji je omjer $(p + n)$ i ukupnog broja vrijednosti (N) u varijablama E i M, zajedno s nasumičnom točnosti $Q_{2,rnd}$ i njihovim rasponom (razlikom) ΔQ_2 analiziranom u [34,35]. Stoga su ti parametri bili početna inspiracija za analize i istraživanja u disertaciji - s ciljem kvantificiranja složenosti varijable. Modelna varijabla M koja ima najbolje slaganje s varijablom E dat će maksimalnu vrijednost parametra točnosti Q_2, a ona modelna varijable koja ima najlošije moguće slaganje s eksperimentalnom varijablom E dat će minimalnu vrijednost. Nadalje, ako napravimo usrednjenje svih vrijednosti parametra točnosti Q_2 izračunanih iz svih parova izmjenjivih varijabli E i M, dobit ćemo prosječnu nasumičnu vrijednosti parametra točnosti, koja odgovara vrijednosti parametra $Q_{2,rnd}$. Taj je važan rezultat dobiven simulacijama provedenim u disertaciji, a pomogao ispraviti raniju definiciju tog parametra iz literature [34,35].

Minimalna, maksimalna i prosječna vrijednost parametra točnosti Q_2 za kompletni skup izmjenjivih varijabli E i M, nazvane su karakterističnim vrijednostima parametra točnosti. Slaganje (podudarnost) vrijednosti binarnih klasifikacijskih izmjenjivih varijabli E i M izražava se i brojnim drugim statističkim parametrima (mjerama kvalitete) [33,78] koji se koriste u širokom području klasifikacijskog modeliranja u raznim znanstvenim područjima od humanističkih znanosti (sociologija, psihologija) preko društvenih znanosti (ekonomija) pa do prirodnih, medicinskih, tehničkih i biotehničkih znanosti. Stoga, prva istraživanja provedena za određivanje karakterističnih vrijednosti parametra točnosti, provedena su i za druge najčešće korištene parametre za iskazivanje podudarnosti (kvalitete slaganja) binarnih izmjenjivih varijabli E i M, poput srednje apsolutne pogreške, standardne pogreške, koeficijenta korelacije [36,37], parametra $F1$ [35-37] i iznimno često korištenog parametra Cohenove kape (κ) [38].

Prvobitni plan bio je izvesti analitički karakteristične vrijednosti parametra točnosti, te potom simulacijama provjeriti ispravnost izvedenih izraza. Za karakteristične vrijednosti drugih parametara $(MAE, s, MCC, F1$ i $\kappa)$ planirano je provesti simulacijska istraživanja i na taj način približno odrediti njihove vrijednosti. Međutim, u radu na izvođenju minimalnih i maksimalnih vrijednosti parametra točnosti uočeno je da je moguće definirati supstitucijske vrijednosti za elemente tablice pogrešaka $(p, n, u$ i $o)$ s pomoću kojih je moguće izvesti i rubne (minimalne i maksimalne) vrijednosti drugih parametara poput $MAE, s, MCC, F1$ i κ. U analogiji s izvodom izraza za nasumičnu točnost $Q_{2,rnd}$ [34], uspjelo se uočiti kako je i tu moguće definirati supstitucijske izraze za elemente tablice pogrešaka $(p, n, u$ i $o)$ za izračun prosječnih nasumičnih vrijednosti drugih parametara, poput $MAE, s, MCC, F1$ i κ [33,35,37,38,73]. Potom, simulacijska istraživanja potvrdila su ispravnost izvedenih izraza za sve karakteristične vrijednosti svih parametara analiziranih u disertaciji.

4.1.4 Izvodi karakterističnih vrijednosti parametara kvalitete modela

Pokazalo se kako je karakteristične vrijednosti parametara Q_2, ΔQ_2, MAE, s, MCC, $F1$ i κ moguće izvesti i algebarskim putem u ovisnosti o udjelu klase 1 (x). U dosadašnjoj praksi, određivanje karakterističnih vrijednosti tih parametara nije rađeno. Supstitucijski izrazi za elemente matrice pogrešaka (p, n, u i o) u ovisnosti o x, koji služe za izračunavanje minimalnih i maksimalnih vrijednosti svih parametara, dani su u *Tablici 2.3* (*Materijali i metode*).

Važno je napomenuti da ti izvedeni izrazi vrijede za izračun podudarnosti (točnosti ili pogrešaka) između izmjenjivih klasifikacijskih varijabli E (eksperimentalna) i M (modelna) s dvije klase (klasa 1 i klasa 0) i s jednakim udjelom klasa [39,40], a takve varijable mogu se nazvati i kao binarne klasifikacijske izmjenjive varijable. To je ekvivalentno analizi podudarnosti vrijednosti između eksperimentalne (E) i modelne (M) varijable (dobivene predviđanjem uravnoteženim modelom, koji predviđa udio klasa identičan varijabli E na kojoj je model razvijen, tj. ugođen), kako je objašnjeno ranije za parametar točnosti Q_2 i za doprinos modela iznad nasumične točnosti [34,35].

Sortiranjem podataka u dva oblika (AA – E i M varijable jednako poredane, AD - E i M varijable nasuprotno poredane) dobivaju se minimalne i maksimalne karakteristične vrijednosti parametara u ovisnosti o udjelu klase 1 (x) koji može poprimati vrijednosti između 0 i 1.

Parametri Q_2, ΔQ_2, MCC, $F1$ i κ podudarnost između binarnih klasifikacijskih varijabli E i M iskazuju u obliku točnosti (tj. vrijednosti su im veće kada je, na istim mjestima, identičan veći broj vrijednosti u obje varijable). Parametri srednja apsolutna pogreška (MAE) i standardna pogreška (s) podudarnost između varijabli E i M iskazuju preko broja odstupanja, te su im vrijednosti veće kada je, na istim mjestima, različit veći broj vrijednosti u obje varijable.

Pojam izmjenjive varijable objašnjene su u poglavlju 2.1.5, i odgovaraju konceptu poznatom u matematičkoj teoriji [39,40]. U poglavlju 3.1.1 (*Tablica 3.3*) dani su izvodi minimalnih i maksimalnih karakterističnih vrijednosti parametara točnosti Q_2 i ΔQ_2 u ovisnosti o x - udjelu klase 1. Pomoću dva oblika sortiranja vrijednosti u varijablama E i M (AA – kada su vrijednosti poredane upareno/jednako i AD – kada su vrijednosti poredane nasuprotno) simulacijama su dobivene minimalne i maksimalne karakteristične vrijednosti parametara u ovisnosti o udjelu klase 1 (x), u rasponu vrijednosti od 1 % do 100 %.

Ideja sortiranja varijabli E i M u poretku AA (za maksimalnu karakterističnu vrijednost Q_2 i ΔQ_2) i u poretku AD (za minimalnu karakterističnu vrijednost Q_2 i ΔQ_2) u postupku izvoda matematičkih izraza za minimalne i maksimalne karakteristične vrijednosti u ovisnosti o x (poglavlje 3.1.1), potekla je iz paralelno rađenih simulacijskih istraživanja. Također je uočeno da su izvodi u poglavlju 3.1.1 morali biti rađeni odvojeno na pod-intervalima vrijednostima $x \leq 1/2$ ili $x \geq 1/2$ (rubna vrijednost $1/2$ može istovremeno pripadati u oba podintervala). Pritom, u dobivanju konačnih izraza koristile su se supstitucije elemenata matrice pogrešaka p, n, u i o, sve izražene u ovisnosti o udjelu klase 1 (x).

Ovisnost minimalnih vrijednosti Q_2 o x je linearnog oblika $1 - 2x$ za $x \leq 1/2$ i $2x - 1$ za $x \geq 1/2$. Maksimalna vrijednost Q_2 u ovisnosti o x uvijek je konstantna i poprima maksimalnu vrijednost jednaku 1. To je posljedica činjenice da su u varijablama E i M izmjenjive i imaju identične udjelele obiju klasa, što znači da postoji jedna permutacija varijable M koja se savršeno poklapa s varijablom E. Ovisnost nasumične točnosti $Q_{2,rnd}$ o x ima oblik parabole ($2x^2 - 2x + 1$) i ta vrijednost ne ovisi o permutacijama varijable u cijelom rasponu x. S obzirom na to da se ΔQ_2

računa kao razlika Q_2 i $Q_{2,rnd}$, njegova karakteristična funkcija je također nelinearna (oblika parabole: $-2x^2 + 2x$), i u potpunosti se podudara s rezultatima dobivenim simulacijama za varijable u rasponu x od 1 % do 100 % u poglavlju 3.1.3. Također, ta se ovisnost podudara s jednadžbom (4.1) za maksimalnu točnost iznad nasumične točnosti koja se može dobiti između predviđanja uravnoteženim modelom (M) i početne eksperimentalne varijable (E).

Rezultati svih izvoda karakterističnih vrijednosti parametara u ovisnosti o udjelu klase 1 (x) nalaze se u *Tablici 4.1*.

Tablica 4.1 Formule maksimalnih i minimalnih karakterističnih vrijednosti parametara korištenih u analizi podudarnosti vrijednosti klasifikacijskih varijabli E i M u ovisnosti o udjelu klase 1 (x)

parametri	poredak varijabli AA [a] $x \in [0,1]$	poredak varijabli AD [a] $x \in [0, \frac{1}{2}]$	poredak varijabli AD [a] $x \in [\frac{1}{2}, 1]$
Q_2	1	$1 - 2x$	$2x - 1$
ΔQ_2	$-2x(x-1)$	$-2x^2$	$-2(x-1)^2$
$Q_{2,rnd}$	$2x^2 - 2x + 1$	$2x^2 - 2x + 1$	$2x^2 - 2x + 1$
MAE	0	$2x$	$2(1-x)$
s	0	$\sqrt{2x}$	$\sqrt{2(1-x)}$
MCC	1	$\frac{x}{x-1}$	$\frac{x-1}{x}$
κ	1	$\frac{x}{x-1}$	$\frac{x-1}{x}$
$F1$	1	0 [b]	$\frac{2x-1}{x}$

[a] AA – slučaj kada su varijable E i M jednako/upareno poredane, AD – slučaj kada su varijable E i M suprotno/obrnuto poredane; [b] parametar $F1$ nije definiran u točki $x = 0$.

Izvodi svih minimalnih i maksimalnih karakterističnih vrijednosti svih ostalih parametara kvalitete (MAE, s, MCC, $F1$ i κ) provedena je u poglavlju 3.4.1, analogno kao i izvodi za odgovarajuće prethodno opisane parametre Q_2 i ΔQ_2.

U *Tablici 4.1*, AA poretkom vrijednosti varijabli E i M dobiva se maksimalna karakteristična vrijednost parametra Q_2, ΔQ_2, *MCC*, *F1* i κ. Obrnutim poretkom vrijednosti u varijablama E i M (AD) dobiva se minimalna vrijednost tih parametara. Za parametre MAE i s koji podudarnost iskazuju zbrajanjem pogrešaka (nepodudarnosti vrijednosti) varijabli E i M, vrijedi obrnuto. Vrijedno je spomenuti kako je osnovni izraz, pa i maksimalne i minimalne karakteristične vrijednosti parametra Cohenove kape κ [38] za izmjenjive varijable E i M identične odgovarajućim izrazima za koeficijent korelacije MCC, što dosad nije spomenuto u literaturi - a to izvedeno je (dokazano) u disertaciji.

Za najvjerojatniju nasumičnu vrijednost Q_2 parametra ($Q_{2,rnd}$) pokazalo se da je potrebna promjena naziva iz najvjerojatnije nasumične vrijednosti parametra točnosti Q_2 iz radova [34] i [35] u prosječnu nasumičnu vrijednost. Simulacijskim analizama dokazalo se da $Q_{2,rnd}$ parametar odgovara aritmetičkoj srednjoj vrijednosti svih Q_2 parametra, te da ta srednja vrijednost nekad ne mora postojati kao stvarna vrijednost parametra. Hoće li srednja vrijednost postojati kao stvarna vrijednost parametra ovisno je o tome je li riječ o parnom ili neparnom broju vrijednosti u

varijablama E i M koje se uspoređuju (preklapaju), tj. o N. Parametar $Q_{2,rnd}$ se pokazalo da ovisi direktno o srednjoj vrijednosti, tj. o udjelu klase 1, pa je time neovisan o poretku podataka u varijabli.

Tako je posljednja karakteristična vrijednost svakog pojedinog preostalog parametra MAE, s, MCC, $F1$ ili κ zapravo procjećna vrijednost svih vrijednosti tog parametra koje se mogu dobiti iz usporedbe (preklapanja) varijable E (uvijek u istom pretku) i neke od permutacija varijable M. Njihovi izvodi dani su u dijelu 3.4.1, osim izvoda za prosječnu vrijednost parametra ΔQ_2 koja je dana u dijelu 3.1.1, i ima konstantnu vrijednost jednaku 0.

U *Tablici 4.2* dane su pojednostavljene supstitucijske vrijednosti za izračun prosječnih vrijednosti parametara za izmjenjive varijable ($u = o$) iskazane u ovisnosti o udjelu klase 1 (x).

Tablica 4.2 Prosječne nasumične vrijednosti parametara iskazana u ovisnosti o udjelu klase 1 u izmjenjivim varijablama E i M

parametri	prosječne nasumične vrijednosti
Q_2	$(2x^2 - 2x + 1)$
ΔQ_2	0
MAE	$2x - 2x^2$
s	$\overline{2x - 2x^2}$
MCC	0
$F1$	x
κ	0

***** N – ukupni broj podataka u varijabli E i u varijabli M; x – udio klase 1 u izmjenjivim varijablama E i M s vrijednostima u rasponu od 0 do 1; U izvodima prosječnih nasumičnih vrijednosti korišteni su ovi supstitucijski izrazi: (1) Nx^2 za p, (2) $N(1-x)^2$ za n, i (3) $Nx\ \ 1-x\ $ za u i o,

Ispravnost izvedenih izraza za prosječne karakteristične vrijednosti dodatnih parametara kvalitete (MAE, s, MCC, $F1$ ili κ) provjerena je i potvrđena simulacijama provedenim u poglavlju 3.4.2 za svaki pojedinačni parametar posebno. Dobiveni rezultati nadmašili su prvobitna očekivanja, i takvi rezultati nisu poznati u literaturi.

U ovom dijelu raspravljeni su izvodi karakterističnih vrijednosti parametara kvalitete. Za svaki parametar to su tri vrijednosti: minimalna, maksimalna, i prosječna vrijednost. Zamisao nastavka istraživanja bila je analizirati njihove vrijednosti, i njihove raspone (tj. razlike između bilo koje dvije karakteristične vrijednosti), te istražiti njihovu korelaciju s entropijom varijable. Parametar ili raspon (jedan ili više njih) koji pokaže visoku korelaciju s entropijom, poslužit će za definiranje kriterija minimalne prihvatljive složenosti (pragove složenosti) klasifikacijske varijable s dva stanja. Klasifikacijska varijabla s dva stanja odgovara u potpunosti indikatorskim varijablama [5,24] uvedenim još u samim početcima u QSAR/QSPR modeliranja [5] te je i danas je u vrlo intenzivnoj uporabi kroz velike skupove (od 100-200 pa do više tisuća) *fingerprints* deskriptora [29,24].

4.1.5 Simulacije karakterističnih vrijednosti parametara kvalitete modela

Prvotno je bilo zamišljeno i planirano da se velika većina karakterističnih vrijednosti većine razmatranih parametara (Q_2, ΔQ_2, MAE, s, MCC, $F1$ i κ) dobije računalnom simulacijom rađenom na velikom broju parova binarnih klasifikacijskih varijabli E i M, od kojih je svaka s $N = 100$

vrijednosti. Simulacije u disertaciji provođene su za parametre Q_2, ΔQ_2, MAE, s, MCC, $F1$ i κ - računane svaki put između parova izmjenjivih varijabli E i M (opisane ranije, $N = 100$) za razne permutacije varijable M - mogu se podijeliti u dvije pod-skupine:

(**1**) Simulacije vezane uz raspodjele vrijednosti pojedinog parametra kvalitete,

(**2**) Simulacije vezane uz analizu ispravnosti izvedenih izraza za karakteristične vrijednosti parametara i kroz njihovu usporedbu s odgovarajućim karakterističnim vrijednostima parametara dobivenih simulacijama.

U prvom **(1)** dijelu simulacija analizirano više vrsta izmjenjivih varijabli s različitim udjelima klasa, a u disertaciji su prikazani rezultati dobiveni samo za izmjenjive varijable E i M s udjelima klase 1 i 0 ($x : (1 - x)$) jednakim 0.5: 0.5 i 0.8:0.2, odnosno, iskazano u postotcima: 50:50 % i 80:20 %. Odabrana su samo ova dva skupa (50:50 % i 80:20 %) od kojih je prvi simetričan slučaj, a drugi značajno nesimetričan, i sasvim su dovoljni za analizu svojstava raspodjele gdje prvi broj predstavlja udio klase 1, a drugi broj udio klase 0 u varijablama E i M. Cilj simulacija bio je provjeriti kako promjena udjela klase 1 (x) utječe na:

a) raspodjelu vrijednosti (i pojedinačne vrijednosti) pojedinog analiziranog parametra,

b) apsolutne minimalne i maksimalne karakteristične vrijednosti pojedinog analiziranog parametra kvalitete,

c) prosječnu karakterističnu vrijednost pojedinog analiziranog parametra kvalitete,

d) raspon karakterističnih vrijednosti pojedinog analiziranog parametra kvalitete.

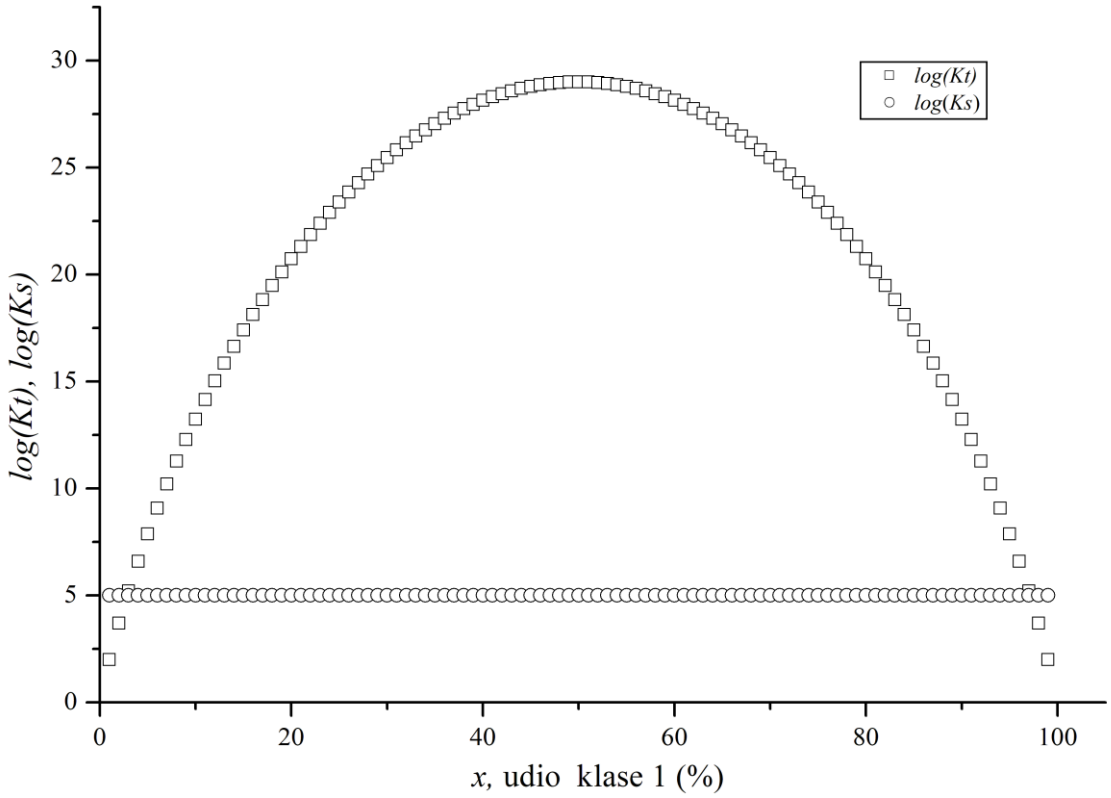
U drugom **(2)** dijelu simulacija analizirano je 99 slučajeva parova izmjenjivih varijabli s udjelom klase 1 (x) u izmjenjivim varijablama E i M između 1 % i 99 % (s korakom 1 %). Pritom, u svakom od 99 pokusa modelna varijabla M permutira se veliki broj puta, i za svaku njenu neidentičnu permutaciju računa se vrijednost parametra podudarnosti (točnosti) između varijabli E i M. To je značajno složeniji dio istraživanja, i sadrži apsolutne simulacijske vrijednosti dobivene:

a) uparenim (AA) i

b) obrnutim (AD) uparivanjem/sortiranjem vrijednosti varijabli E i M.

Poredak AA daje maksimalne karakteristične vrijednosti parametara za Q_2, ΔQ_2, MCC, $F1$ i κ, te minimalne za parametre (pogreške) MAE i s, koji podudarnost iskazuju kao pogrešku. Poredak AD daje minimalne karakteristične vrijednosti parametara za Q_2, ΔQ_2, MCC, $F1$ i κ, te maksimalne za parametre (pogreške) MAE i s.

Za varijablu M s $N = 100$ vrijednosti (od kojih njih X ima vrijednost 1, i $N - X$ vrijednost 0) broj mogućih permuatcija vrijednosti jako je velik - i on iznosi $100!$ [$X!$ $100 - X!$]. Zbog toga je u svim pokusima za varijable s udjelom klase 1 (x) u izmjenjivim varijablama E i M između 1 % i 99 % automatski načinjeno 10^5 permutacija. Na *Slici 4.1* dan je prikaz logaritma teorijskog broja mogućih neidentičnih permutacija varijable M ($log(K_t)$) u ovisnosti o udjelu klase 1 (x) u postotnom rasponu 1 % i 99 %.



Slika 4.1 Obrađeni broj parova varijabli E i M u simulacijama u odnosu na najveći mogući broj permutacija varijable M

Također, na grafu prikazanom na *Slici 4.1* dan je i logaritam broja permutacija proveden u simulacijama ($log(K_s)$) za svaki primjer varijable (za svaki x, %), i on je stalan i jednak 5 (= $log(10^5)$. Vidimo da samo za male i za velike udjele klase 1 (x), broj permutacija u simulacijama blizak je teorijskim. Inače, on je jako mali dio mogućih teorijskih permutacija varijable M (*Slika 4.1*) u paru izmjenjivih varijabli E i M, koje su podloga svih simulacija u disertaciji. Na temelju tako velike razlike, moglo bi se očekivati kako je za vjerodostojnije simulacijske rezultate potrebno povećati broj permutacija varijable M. Međutim, rezultati su pokazali da to uopće nije potrebno.

Simulacije provedene za karakteristične vrijednosti parametra Q_2 za raspodjele simulacijskih vrijednosti varijabli 50:50 % i 80:20 % (*Slike 3.1* i *3.2*) pokazuju asimetriju i suženje raspona maksimalnih i minimalnih karakterističnih vrijednosti. Također, jasno se vidi pomak srednje vrijednosti tih parametara s 50 % na 68 % kod varijable s udjelima klasa 80:20 % u odnosu na varijablu 50:50 %. Slično se pokazuje za raspone između maksimalnih i minimalnih karakterističnih vrijednosti koji postaju manji kod varijabli 80:20 % (40 %) u odnosu na raspon 100 % kod varijabli E i M s udjelima klasa 50:50 %. To ukazuje na činjenicu da varijabla s udjelima klasa 80:20 % ima značajno manju složenost, što smo i očekivali. Posve slični zaključci dobiveni su za analize provedene za parametar stvarne točnosti iznad nasumične ΔQ_2 (*Slike 3.3* i *3.4*).

Nadalje, simulacije (drugi dio) vezane uz analizu ispravnosti izvedenih izraza za karakteristične vrijednosti Q_2 i ΔQ_2 iskazane preko udjela klase 1 (x) između 1 % i 99 % (s korakom 1 %) u izmjenjivim varijablama E i M pokazale su ispravnost svih izvedenih izraza zbirno prikazanih u *Tablicama 4.1* i *4.2*. Ta se ispravnost očituje i identičnošću funkcionalnih ovisnosti na *Slikama 3.5* i *3.7* u ovisnosti o x i izvedenih formula. Nadalje, simulacijske vrijednosti parametara Q_2 i ΔQ_2 nikad

nisu postigle veću vrijednost od maksimalnih apsolutnih vrijednosti, niti nižu vrijednost od apsolutnih minimalnih vrijednosti parametara.

Potpuno analogni zaključci dobiveni su za ostale simulacijske analize karakterističnih vrijednosti parametara MAE, s, MCC, $F1$ i κ i njihovih raspodjela (poglavlje 3.4.2). Nadalje, ispravnost izvoda njihovih karakterističnih vrijednosti tih parametara (prikazanih u *Tablicama 4.1* i *4.2*) i raspona njihovih karakterističnih vrijednosti potvrđene su u rasponu udjela klase 1 (x) od 1 % do 99 % u poglavlju 3.4.3. Ti su rezultati redom prikazani na *Slikama 3.20* do *3.27*.

Iz tih rezultata izdvajamo jedno korisno saznanje u vezi parametra srednje apsolutne pogreške. Kako je MAE parametar (prosječna apsolutna pogreška) vrlo sličan Q_2 parametru (komplementarna formula) tako su i njegove karakteristične vrijednosti parametara slične, tj. riječ je o pravcima (*Slika 3.20* i *Slika 3.5*). Jedna razlika je da MAE iskazuje podudarnost preko pogreške – tj. broja nepodudarnih vrijednosti, a Q_2 preko broj točnih vrijednosti u varijablama E i M.

Apsolutne maksimalne vrijednosti i apsolutni rasponi karakterističnih vrijednosti parametra MAE, s, MCC, $F1$ i κ u ovisnosti o udjelu x uvijek su veće od simulacijskih, što je dodatna potvrda točnosti formula izvedenih u *Tablicama 4.1* i *4.2*. Isto pravilo vrijedi za minimalne apsolutne vrijednosti koje su uvijek manje od simulacijskih. Usporedbom izvedenih formula karakterističnih vrijednosti u ovisnosti o udjelu x s numeričkim podacima iz simulacija, potvrđena je točnost izvoda. Kod provjera ispravnosti izvedenih izraza za prosječne nasumične vrijednosti parametara (kao treće karakteristične vrijednosti) provedeno je i više simulacija na varijablama s malim brojem vrijednosti (npr. sa samo 4 ili 5 vrijednosti) pri čemu su analizirani svi parovi izmjenjivih varijabli E i M (tj. sve permutacije varijable M). I te su simulacije potvrdile kako su prosječne vrijednosti parametara kvalitete za izmjenjive varijable dane točno izrazima u *Tablici 4.2*.

4.2 Entropija varijable i njena korelacija s karakterističnim vrijednostima parametara kvalitete modela

4.2.1 Entropija varijable

U ovom radu je prvi put uveden pojam entropije kao mjere složenosti varijable u QSAR/QSPR modelima [24]. Korištena je prilagođena Boltzmannova formula za entropiju ($logW$) dana jednadžbom (3.20) kao preciznija mjera složenosti. Naime, u literaturi se u pravilu koristi Shannonov oblik entropije [70] koja koristi Stirlingovu aproksimaciju [71] (jedn. (3.22)) u izračunu faktorijela. Taj izraz spominje se i koristi u raznim kontekstima i analizama i kvantificiranju složenosti molekularnih struktura, te u QSAR/QSPR modeliranju [24]. Međutim, taj izraz (Shannonov oblik entropije) nije primjenjiv na varijable s malim brojem vrijednosti - već na one s manje od 100 vrijednosti). Preciznije, pogreška Stirlingove aproksimacije u računanju faktorijela postaje zanemariva tek nakon $N = 150$. Iako je moguće poopćiti izraz za izračun entropije na više klasa, u disertaciji su provedena istraživanja na najjednostavnijem konceptu koji se tiče složenosti pojedinog molekularnog deskriptora koji ima samo dvije vrijednosti (0 i 1) – koji se u literaturi još naziva i indikatorski deskriptori [5,24].

4.2.2 Korelacija entropije varijable s karakterističnim vrijednostima parametara kvalitete modela i njihovim rasponima

Nenormalizirana formula entropije ($logW$) uspoređena je sa svim ciljanim parametrima ($Q_2, \Delta Q_2$ i $Q_{2,rnd}$) čija je vrijednost ovisna o x. Parametar $Q_{2,rnd}$ pokazao je najveću korelaciju s $logW$, a samim time i parametar $\Delta Q_{2,max}$ koji je ovisan o $Q_{2,rnd}$, a i njegova vrijednost korelacije s entropijom je po apsolutnom iznosu jednaka, ali suprotnoga predznaka, što je i razumljivo i što ne mijenja smisao zaključaka. Najviša korelacija $\Delta Q_{2,max}$ s entropijom varijable ima jasnu interpretaciju – tj. da je složenija varijabla kod koje maksimalni mogući doprinos modela maksimalan. To ujedno i potvrđuje početnu pretpostavku u disertaciji.

Osim triju osnovnih parametara (Q_2, ΔQ_2 i $Q_{2,rnd}$) u *Tablici 3.9* parametar $logW$ koreliran je i s parametrima MAE, s, MCC i $F1$. Korelirane su kako minimalne i maksimalne karakteristične vrijednosti, tako i nasumične vrijednosti, te razlika svih karakterističnih vrijednosti. U toj analizi korelacije s $logW$ parametar MAE ($R = 0.997$, za cijeli raspon varijabli od 0 % do 50 %, *Tablica 3.9*) pokazao se podjednako dobar kao i prosjek Q_2 parametra – odnosno prosječna nasumična točnost $Q_{2,rnd}$. Usporedbom karakterističnih vrijednosti i raspona, parametar s (standardna pogreška) pokazao je nešto nižu korelaciju s entropijom $logW$ ($R = 0.990$, *Tablica 3.9*) nego parametar MAE (prosječne nasumične pogreška), iako je s daleko češće korišteni parametar u računanju pogreške u svim zanstvenim područjima. Inače, vrijedno je istaknuti i da se parametri pogreške često rabe u literaturi kod analize podudarnosti (poklapanja ili odstupanja) između para kontinuiranih varijabli, daleko češće nego koeficijenta korelacije. Međutim, u analizi binarnih klasifikacijskih varijabli, skoro se ne rabe parametri pogreške. Primjer za to je iznimno često citirani pregledni rad Powersa [33], u kojem pogreške (MAE i s) nisu ozbiljno razmatrane kao važni parametri u procjeni točnosti binarnih klasifikacijskih modela. Ovi rezultati upućuju na to da bi trebalo predložiti promjenu takve prakse, što će biti dodatno istraženo u budućnosti. Nadalje, ove korelaciju upućuju na to da kod binarnih klasifikacijskih varijabli postoji ekvivalencija između maksimalnog doprinosa modela iznad nasumične točnosti ($\Delta Q_{2,max}$) i srednje apsolutne pogreške (MAE), i da su ti cparametri u najboljem slaganju s entropijom varijabli koje se uspoređuju .

4.2.3 Normalizirana entropija i normalizirana točnost

S ciljem praktične primjene koncepta složenosti (iskazanog preko entropije varijable) u postupku eliminacije deskriptora niske značajnosti iz QSAR/QSPR modela [22,24], uveden je koncept normalizirane entropije $logW_{norm}$ jednadžbom (3.25) (poglavlje 3.2). Normalizacija je ovisna o broju elemenata (vrijednosti, instanci, molekula) u varijabli koja se analizira (N), i ona se provodi za svaku varijablu, odnosno za skup varijabli/deskriptora izračunanih za određeni skup molekula. Faktor normalizacije je najsloženija moguća varijabla s N vrijednosti, te se složenost pojedine varijable iskazuje u postotcima maksimalne moguće složenosti. Kako bi se moglo lakše odlučiti o prihvatljivoj složenosti, definirane su razine (pragovi) složenosti od 1 %, 2.5 %, 5 %, 7.5 % i 10 %.

Analogno je postupljeno s maksimalnim mogućim doprinosom modela iznad nasumične točnosti $\Delta Q_{2,max}$, (odnosno ΔQ_2), i normalizirana vrijednost tog parametra označena je s $\Delta Q_{2,max,norm}$. Faktor normalizacije u tom slučaju je maksimalni mogući doprinos modela iznad nasumične točnosti dan izrazom (4.2) i on iznosi $1/2$ (u pravilu). Taj će broj biti malo različit (veći) od $1/2$ samo onda kada je broj vrijednosti (tj. molekula u skupu) N u varijablama E odnosno M

neparan. Taj je parametar iskorišten također kao mjera složenosti, jednako kao i normalizirana entropija (Rezultati - formula (3.24)), a njihova praktična primjena omogućena je, mrežnim poslužiteljem razvijenim u disertaciji [76].

Razine značajnosti prikazane u rezultatima u disertaciji čine se kao jednostavna i prikladna mjera složenosti, daleko jednostavnija i bliža izvornom konceptu entropije (permutacijske entropije, vezano s brojem mogućih permutacija varijable) temeljenom na particijskoj funkciji nego je to slučaj u literaturi (npr. u radu [85]). Iako se prema naslovu rada [85] čini da je riječ o vrlo sličnom konceptu, u stvarnosti je to posve različit pristup koji nije vezan s entropijom – riječ entropija uopće se ne spominje u tom radu). U budućnosti se biti potrebno dodatno razraditi uvjete i način razina značajnosti kod izračuna složenosti varijable i njihove primjene u postupku eliminacije varijabli niske varijabilnosti i niske značajnosti (npr. konstantne ili skoro konstantne varijable/deskriptore) [24], te tako optimirati i njihovu daljnju praktičnu uporabu.

Imajući na umu da podskup varijabli koji će se izabrati u konačnom multivarijatnom modelu treba imati visoku korelaciju s eksperimentalnom varijablom (izmjerenim svojstvom/aktivnošću molekula) (y) i nisku korelaciju s ostalim varijablama (x) [86,87], preporuka je da se u algoritme za izbor varijabli uključi i jedna od mjera složenosti podataka. Na taj način, u modelu bi u svakom koraku odabira uvijek ostajala varijabla dovoljne (prihvatljive) razine složenosti. Taj je problem osobito izražen danas kada sve više binarnih klasifikacijskih podataka koji se mjere i modeliraju neuravnoteženo [88,89], tj. posjeduju daleko veći broj vrijednosti koji pripadaju jednoj klasi, dok je broj vrijednosti koje pripadaju drugoj klasi jako mali. U tom slučaju pitanje složenosti varijabli koje se razmatraju u modeliranju i pitanje prosječne nasumične točnosti daleko je važnije. Naime, kod velike neuravnoteženosti vrijednosti klasifikacijske varijable y koja predstavlja aktivnost koja se modelira, sama y varijabla ima daleko nižu složenost od maksimalno moguće. To ima posljedice i na izbor deskriptora u model koji se razvija ugađanjem vrijednosti jedne takve varijable.

Između $\Delta Q_{2,max,norm}$ i $logW_{norm}$ postoji visoka korelacija (*Slika 3.10*), a vidimo da je uvijek normirana entropija nešto veća (za isti X) u cijelom rasponu broja vrijednosti elemenata (instanci, molekula) klase 1 (X) u rasponu 0 do 100. Normirana vrijednost entropije pokazala se manje strogom mjerom složenosti, i po tom normiranom parametru složenost varijable ranije postiže/prelazi viši prag složenosti, za isti udio klase 1 (x). Oba normirana parametra koriste se i u procjeni složenosti kontinuiranih varijabli, tako što se te varijable dihotomiziraju u odnosu na srednju vrijednost svake pojedine varijable.

4.3 Primjena rezultata

4.3.1. Primjena na skupovima molekularnih deskriptora

Normirani parametri $\Delta Q_{2,max,norm}$ i $logW_{norm}$ koji su izdvojeni kao najbolji za procjenu složenosti varijabli implementirani su u mrežnom poslužitelju „*Classification variable complexity parameter estimator*“ [76]. Analizirana je složenost četiri skupa deskriptora iz literature u dijelu 3.6.2: *Huuskonen30* i *Huuskonen58* [66] s 30 odnosno 58 deskriptora za 884 molekule, te skupovi *taksani* i *pacitakseli* [67] svaki s jednom y-varijablom (biološka aktivnost) i tri deskriptora. (*Tablice 3.10 i 3.13*). Primijećeno je da je u skupovima *Huuskonen30* i *Huuskonen58* 5 od 30 (*Tablica 3.10*) odnosno 15 od 58 deskriptora (*Tablica 3.11*) imalo složenost prema parametru $\Delta Q_{2,max,norm}$ ispod praga od 10 %, te ih je moguće isključiti na temelju malog udjela jedne od klasa. Za ta dva skupa

prema normaliziranoj entropiji $logW_{norm}$složenost manju od 10 % ima 3 od 30 odnosno 10 od 58 deskriptora. Analiza manjih skupova s 33 taksana (*Tablica 3.13*) i 22 pacitaksela (*Tablica 3.15*) pokazala je da svi deskriptori i y-varijable prelaze prag složenosti od 10 % po oba normirana parametra složenosti.

Analiza skupova deskriptora izračunanih aplikacijama razvijenim ovim doktorskim radom: (1) ProtSeqAnalyzer koja za proteinske sekvence kreira listu motiva i njihovih frekvencija [82] i (2) „*Zagreb indices and their modifications – CALCULATOR"* za računanje skupa topoloških deskriptora [53] primijenjena je u analizi skupa 568 antimikrobnih peptida. Pokazalo se da od ukupno 1941 deskriptora njih 293 ne prelazi prag složenosti od 10 % prema parametru $logW_{norm}$, te 394 prema normaliziranoj maksimalnoj mogućoj točnosti iznad nasumične ($\Delta Q_{2,max,norm}$).

Analiza na skupovima pokazala je potrebu za nastavkom istraživanja s ciljem postrožavanja kriterija složenosti osobito kod varijabli (1) s malim brojem vrijednosti i (2) koje imaju mali broj vrijednosti koje pripadaju jednoj klasi (tj. neravnoteža podataka). U slučaju uključenja u model deskriptora koji ima samo jednu vrijednost jednaku 1 (a preostalih $N-1$ ima vrijednosti 0), takav bi deskriptor unosio u model minimalno jednu konstantu koja bi ugađala (*engl*. fit) samo tu jednu vrijednost koja je jednaka 1. Stoga, u tom slučaju uključivanje jedne takve varijable/deskriptora u model ne bi doprinosilo sposobnosti generalizacije modela – što je cilj svakog modeliranja. Naime, ukoliko bismo modelom s N koeficijenata (optimiranih parametara) od kojih svaki dolazi uz jedan od N deskriptora, imali bismo sustav od N jednadžbi s N nepoznanica. Poznato je iz osnova modeliranja (i matematičke statistike) da takav model ne bi imao nikakvu sposobnost generalizacije. Stoga, logično je uzeti da deskriptor mora imati minimalno dvije vrijednosti (za dvije molekule u skupu) u manjinskoj klasi (npr. klasi 1), a preostalih $N-2$ vrijednosti u dugoj klasi (klasi 0). Prema tom kriteriju, u ovom slučaju deskriptora za 568 peptida, iz modeliranja bi se isključilo 25 od 45 deskriptora a minimalna složenost takve varijable je $\geq$ 2.5 % prema normaliziranoj entropiji te $\geq$ 1 % prema normaliziranoj stvarnoj točnosti modela. Međutim, ako to pravilo primijenimo na manje skupove (npr. $N = 22$, kao u najmanjem analiziranom skupu iz literature [69]), onda bi minimalna prihvatljiva razina složenosti deskriptora sa samo dvije ($X = 2$) vrijednosti u klasi 1 i 20 u klasi 0 za takav skup bila 33 % prema parametru $\Delta Q_{2,max,norm}$, odnosno 40 % prema normaliziranoj entropiji $logW_{norm}$, što je puno više od praga složenosti 10 %.

4.3.2. Primjena u rangiranju modela

U rangiranju modela na prediktivnim natjecanjima [90,91,92] temeljenim na predviđanjima i kontinuiranih i binarnih klasifikacijskih vrijednosti, vrlo često se koristi kao parametar i koeficijent korelacije. Zbog njegovog nedostataka (osjetljivost na konstantan pomak i sistematsku pogrešku) kod vanjskih (test) skupova te zbog preoptimističnih rezultata kod prevelikih ili premalenih vanjskih (test) skupova predloženo je da ga se ne koristi u te svrhe [35]. Predloženo je korištenje standardne pogreške (s) za usporedbu kvalitete modela u predviđanju na vanjskom (test) skupu.

Izraz za izračun MCC u slučaju binarnih klasifikacijskih varijabli dao je Matthews u radu [73] i taj je rad jako često citiran u znanstvenoj literaturi. U radu [74] autori su pokazali da MCC nije dobra mjera kvalitete modela u slučaju binarnih varijabli u kojima prevladava jedna klasa (neuravnotežene raspodjele vrijednosti varijabli koje se koreliraju). Sličan nedostatak vezan s neosjetljivošću parametra MCC na sustavnu pogrešku u predviđanju, tj. na konstantni pomak predviđenih vrijednosti, a taj nedostatak ilustriran je na primjerima korelacije kontinuiranih varijabli [35]. Međutim, posve je analogno ponašanje MCC i u slučaju binarnih varijabli. Stoga, slabija

korelacija entropije varijable s MCC (u rasponu udjela klase 1 (x) u varijabli od 0 % do 50 %) u odnosu na korelacije s parametrima točnosti Q_2, MAE) i standardne pogreške (s) (*Tablica 3.9*), ukazuje na to da je MCC manje prikladan u primjeni na procjeni podudarnosti varijabli u širokom rasponu udjela klase 1 (x). Asimetričnost raspodjele MCC jako ovisi o omjeru udjela klase 1 i klase 0. Ta osobina raspodjele utječe na osjetljivost parametra koja nije jednaka u cijelom području mjerenja [74]. Iako MCC parametar pokazuje veliku osjetljivost o nesimetriji raspodjele vrijednosti (u klasi 1 i 0) varijabli između kojih se računa, drugi autori ga smatraju najboljim izborom u predviđanjima kod neuravnoteženih skupova [37,81].

Svi dosadašnji parametri pokazali su simetričnost ovisnosti karakterističnih vrijednosti u odnosu na udio $x = 1/2$ (npr. *Slika 3.24* za MCC), odnosno, smanjenjem udjela za određenu vrijednost dobiva se ista karakteristična vrijednost koja bi se dobila povećanjem udjela x za određenu vrijednost. Jedini nesimetričan parametar u odnosu na udio $x = 1/2$ je $F1$ parametar koji ovisi samo o elementima tablice pogrešaka p i u, pa je zbog toga za $x \geq 1/2$ konstantna njegova karakteristična vrijednost parametra (i minimalna i maksimalna) (*Slika 3.26*). U intervalu $x \leq 1/2$, minimalna karakteristična vrijednost parametra $F1$ pokazuje jasnu funkcionalnu ovisnost ($2 - 1/x$). Problem parametra $F1$ su i rubne točke (npr. $x = 0$) gdje parametar nije definiran (može se izračunati uporabom limesa, što je potrebno posebno predvidjeti u programskom kodu).

U slučaju klasifikacijskih modela razlika između stvarne točnosti i prosječne nasumične srednje vrijednosti (ΔQ_2) pokazuje vrlo dobre karakteristike u usporedbi s parametrima koji se trenutačno prevladavajuće koriste u rangiranju modela prema prediktivnoj kvaliteti [48,79,80], i pri tome ima i jednostavnu interpretaciju [35]. Analizirane su i vrijednosti elemenata tablice pogrešaka gdje se pokazalo da bolje rangirani ΔQ_2 modeli imaju više točnih predviđanja manjinske klase p i više uravnotežen omjer pogrešaka o i u od odgovarajućih najboljih modela rangiranih prema Q_2, MCC ili $F1$. Simetrija pogrešaka o i u korisna je karakteristika validacijskog parametra u literaturi (Baldi i dr. [55]). Dodatno, parametar ΔQ_2 uvijek je definiran za svaki skup vrijednosti tablice pogrešaka, te ima linearno proporcionalnu vrijednost u odnosu na promjene vrijednosti u supstitucijskoj tablici.

U slučaju uravnoteženih modela, tj. kod usporedbe podudarnosti izmjenjivih varijabli, parametar Cohenova kapa parametar (κ) [38] ima identičan osnovni izraz onomu za MCC, pa time i iste karakteristične vrijednosti. To znači, da se kod optimalnih i ispravno optimiranih modela, a to su uravnoteženi modeli κima ste vrijednosti kao i parametar MCC. Taj rezultat nije dosad objavljen u literaturi.

4.4. Poopćenje rezultata i njihova primjena na drugim problemima

Osim izmjenjivih varijabli [39,40] koje su primarno istraživane u ovoj disertaciji, izvedeni su supstitucijski izrazi koji omogućuju proširenje područja primjene rezultata iz disertacije na usporedbe podudarnosti (točnosti, poklapanja) općenitih parova binarnih klasifikacijskih varijabli. To znači da varijabla E (eksperimentalna) i M (modelna) ne moraju imati isti omjer (udio) klasa 1 i 0. U tom općenitom slučaju, u varijabli E taj omjer se i dalje označava s x, dok se u varijabli M on označava s y, a karakteristične vrijednosti parametara dane su u *Prilogu 3.37*).

Općenita tablica supstitucija (*Tablica 3.19*) može se koristiti za izračun prosječnih vrijednosti parametra. Njome se elementi tablice pogrešaka p, n, u i o zamijenjuju odgovarajućim izrazima kojim se ugrađuju u osnovne formule za izračun razmatranih parametara: MCC, MAE, S, $F1$, κ.

Koristeći supstitucijske izraze (*Tablice 3.19*) izračunane su prosječne nasumične vrijednosti, za sve razmatrane parametre iz ovog rada (*Tablica 3.20*). To je važan rezultata koji može analogno biti primijenjen po potrebi i na drugim parametrima. Prosječne nasumične vrijednosti moguće je prikazati uz pomoć formula za udjele kako izmjenjivih tako i svih ostalih varijabli, što još nije učinjeno ali se planira napraviti u nastavku istraživanja iz disertacije.

Tablica supstitucijskih izraza za elemente tablice pogrešaka p, n, u i o u ovisnosti o udjelu klase 1 (x) (*Tablica 2.3*) primijenjene u određivanju minimalnih i maksimalnih vrijednosti parametara Q_2, $Q_{2,rnd}$, ΔQ_2, s, MAE, $F1$ i κ otvara također mogućnost primjene i na brojne parametre u klasifikacijskom QSAR/QSPR modeliranju [24], ali i u strojnom učenju s primjenama u raznim područjima [33].

Istraživanja provedena u disertaciji otvorila su i novi smjer istraživanja kvalitete modela uvodeći u razmatranje nasumičnu točnost te minimalne i maksimalne vrijednosti bilo kojeg parametra (mjere) kvalitete klasifikacijskog modela s dva stanja. Pomoću izvedenih izraza koji su provjereni simulacijskim istraživanjima, dobiveni rezultati mogu se poopćiti na bilo koje modele (ne samo one uravnotežene), a glavni rezultati mogu se poopćiti na klasifikacijske varijable i modele s tri ili više klasa, što će biti predmet istraživanja u budućnosti.

5. ZAKLJUČAK

Istraživanje u disertaciji odnosi se na razvoj parametara za analizu složenosti klasifikacijskih varijabli s dva stanja (0 i 1) koje imaju N vrijednosti. Takve varijable nazivamo i binarnim klasifikacijskim varijablama. Promatra se par varijabli $(E,\ M)$, od kojih je prva uvijek ista i odgovara varijabli u stvarnom poretku, te ju nazivamo eksperimentalna varijabla (E). Takva varijabla E može predstavlja neko binarno klasifikacijsko svojstvo ili biološku aktivnost skupa molekula (npr. „1“ = toksična molekula; „0“ = netoksična molekula), a može predstavljati i neki molekularni deskriptor (npr. „1“ = molekula ima -OH skupinu; „0“ molekula nema –OH skupinu).

Druga varijabla u paru naziva se modelna varijabla (M), i ona predstavlja neku od permutacija početne varijable E. Varijablu M u takvom paru (E,M) možemo promatrati i kao varijablu predviđenu nekim modelom. Ukoliko zamislimo da je model savršen u predviđanju, imat ćemo par identičnih varijabli $(E,\ E)$, pa će se promatrati i takav par varijabli. Ako zamislimo sve moguće neidentične modele (varijable M) za predviđanja nekog eksperimentalnog svojstava, onda će njih biti onoliko koliko je neidentičnih permutacija varijable E, uključujući još u tim analizama i samu varijablu E koja odgovara savršenom modelu.

Tijekom rada na interpretaciji dobivenih rezultata u disertaciji, uočeno je da su opisani parovi binarnih klasifikacijskih varijabli, tj. varijabla E s N vrijednosti u stvarnom poretku i modelna varijabla M, koja je jedna od permutacija varijable E, u matematičkoj literaturi nazivaju izmjenjivim varijablama [39,40]). Takve varijable imaju identične raspodjele vrijednosti, tj. podjednak broj vrijednosti u klasi 1 (X), pa time i onih u klasi 0 $(N-X)$.

Analizom podudarnosti (preklapanja) parova varijabli E i M dobivaju se elementi tablice pogrešaka p,n,u i o. na temelju kojih se računaju u disertaciji vrijednosti svih parametara kvalitete podudarnosti (točnosti, poklapanja) klasifikacijskih varijabli. Najjednostavniji parametar za iskazivanje podudarnosti dviju binarnih klasifikacijskih varijabli s dva stanja naziva se točnost (ili postotna točnost), i označava se s Q_2 ('Q' je od engleske riječi *Quality*, dok je '2' oznaka da se radi o binarnim varijablama s dva stanja). Taj parametar koristi se za iskazivanje postotne točnosti predviđanja dobivenog s modelom (varijabla *M*) u odnosu na stvarnu eksperimentalnu varijablu *E*. U istraživanju složenosti varijabli u disertaciji prvi cilj bio je teorijskim razmatranjima (algebarskim putem) pronaći karakteristične vrijednosti parametara točnosti Q_2. Pritom, tri su karakteristične vrijednosti:(**1**) minimalna $Q_{2,min}$ i (**2**) maksimalna $Q_{2,max}$ moguća podudarnost među svim parovima (i) izmjenjivih varijabli (odnosno točnost predviđanja varijable *E* modelnom varijablom *M*), i (**3**) najvjerojatnija/prosječna nasumična vrijednost $Q_{2,rnd}$. Sve te karakteristične vrijednosti su u rasponu od 0 do 1 (ili u postotcima od 0 do 100 %).

Izvodi izraza za minimalnu vrijednost Q_2 $(Q_{2,min})$ napravljeni su posebno za dva pod-intervala udjela klase 1 (x) u varijablama E i M za koji vrijedi $x \in [0,1]$), tj. za $x \leq 1/2$ i za $x \geq 1/2$. Maksimalna vrijednost parametra točnosti (Q_2) koji mjeri najbolje moguće slaganje između parova izmjenjivih varijabli E i M uvijek je jednaka 1. Računanjem razlika između parova karakterističnih vrijednosti parametara dobiveni su i analizirani njihovi rasponi (razlike). To znači, računale su se razlike između maksimalne i prosječne vrijednosti $(\Delta Q_{2,max})$ te između minimalne i prosječne vrijednosti $(\Delta Q_{2,min})$, itd. dok se ne iscrpe svi mogući parovi.

Točnost izvedenih izraza temeljenih na veličinama *p, n, u* i *o,* dobivenih teorijskim razmatranjima, uspoređena je i potvrđena pomoću odgovarajućih rezultata dobivenih simulacijama. U prvobitnom planu istraživanja u disertaciji očekivano je bilo dobiti točne algebarske izraze za $Q_{2,max}$ i $Q_{2,min}$. Izraz za $Q_{2,rnd}$ koji je poznat u literaturi, a ponovno analiziran i razmatran u nedavno objavljenom radu [34]. Pošto se nije nazirala mogućnost dobivanja točnih algebarskih izraza za ostale parametre točnosti modela, planirano je dobiti simulacijama većinu karakterističnih vrijednosti svih parametra točnosti modela Takvi točni izrazi do danas nisu izvedeni/poznati u znanstvenoj literaturi - a uspjelo ih se po prvi puta izvesti u sklopu istraživanja provedenih tijekom izrade ove disertacije.

U simulacijskim analizama, $Q_{2,min}$ (kao i minimalna/najmanja moguća vrijednost bilo kojeg parametra točnosti koji računa podudarnost/točnost varijabli E i M) dobije se kada se prva varijabla u paru E sortira (poslaže) uzlazno, a druga varijabla u paru (M) silazno, te se izračuna točnost (podudarnost) takvoga para varijabli. Slično, $Q_{2,max}$ (kao i maksimalna/najveća moguća vrijednost bilo kojeg parametra točnosti koji računa podudarnost/točnost varijabli E i M) dobivao se u simulacijama usporedbom podudarnosti (točnosti) između varijabli E i M kad su obje varijable sortirane (poslagane) uzlazno (ili, što je isto, kad su obje varijable sortirane silazno). Zbirni rezultati dobiveni za minimalne i maksimalne karakteristične vrijednosti svih parametara razmatranih u disertaciji u ovisnosti o udjelu klase 1 (x) objedinjeni su u *Tablici 4.1*.

Analizom izraza za karakterističnu nasumičnu vrijednosti parametra ΔQ_2 (označenu kao u $\Delta Q_{2,rnd}$ [34]) uočena je pravilnost zamjena veličina *p, n, u* i *o* kada se iz izraza za te parametre žele izvesti izrazi za njihovu nasumičnu vrijednost. Tako je definirana supstitucijska tablica za dobivanje prosječnih nasumičnih vrijednosti parametara kvalitete modela s izrazima (funkcijama temeljenim na *p, n, u* i *o*) za svaku od veličina *p, n, u* i *o* koji se uvrštavaju u izvorne formule parametara kvalitete modela kako bi se dobila najvjerojatnija (tj. prosječna) nasumična vrijednost.

Prosječna nasumična vrijednost parametra MCC izvedena pomoću supstitucijskih vrijednosti za *p, n, u* i *o* za izračun nasumičnih vrijednosti (Rezultati - *Tablica 3.19*) iznosi 0. Takav rezultat dobiven je i potvrđen simulacijama. To je potvrđeno na primjeru parametra MCC da formule dobivene s pomoću supstitucijskih vrijednosti iz *Tablice* 3.19 odgovaraju prosječnoj (a ne najvjerojatnijoj nasumičnoj) vrijednosti parametara (mjera kvalitete modela) u općenitom slučaju, tj. ne samo za uravnotežene klasifikacijske varijable i s dva stanja, nego za općenite (bilo kakve) klasifikacijske varijable E i M. Na primjeru parametra MCC to je dodatno jasno , jer za varijable M koje imaju različiti broj elemenata klase 1 i klase 2 nikad nije moguće dobiti $MCC = 0$.

Supstitucijski izrazi za elemente tablice pogrešaka *p, n, u* i *o* (*Tablica 3.19*) za dobivanje prosječnih nasumičnih vrijednosti parametara kvalitete modela iskazani su u ovisnosti o udjelu klase 1 (x) pri čemu je $x = (p + u)/N$ (*Tablica 4.2*). U simulacijama kod svih parametara vidljivo je da promjena udjela klasa $x = (p + u)/N$ mijenja i iznose karakteristične vrijednosti svih parametara, pa time i njihove raspone. Uočava se smanjenje raspona svih parametra za vrijednosti - a koje su više udaljene od vrijednosti $x = 1/2$.

Kod prikazivanja histograma, prikazani su samo rezultati simulacija za udjele klase 1 u varijablama s jednakim 50:50 % udjelom i s udjelom 80:20 %. Na taj se način mogao ilustrirati utjecaj neuravnoteženosti klasa na karakteristične vrijednosti parametara kvalitete i njihove raspone. Nakon toga napravljena je i opsežnija simulacija, gdje se udio klase 1 (x) kreće od 1 do 99 %. Osim toga testirani su izvodi karakterističnih funkcija parametara u ovisnosti o x -u, koje su se pokazale točne za sve izvedene karakteristične vrijednosti – koje su zbirno dane u *Tablicama 4.1* i *4.2*.

Kao mjera složenosti varijable uzeta je prilagođena formula Boltzmannove entropije. Svaka od karakterističnih vrijednosti parametara i njihovi rasponi izraženi u ovisnosti o x (udio klase 1) korelirani su s entropijom $logW$. Najbolja dobivena korelativna veza dobivena je s vrijednostima parametra ΔQ_2. Stoga su entropija $logW$ i parametar ΔQ_2 te normalizirana entropija i normalizirani parametar ΔQ_2 odabrani kao najbolja mjera složenosti klasifikacijskih varijabli binarnih varijabli. Normaliziranje se vrši tako da se vrijednost $logW$ parametra podijeli s maksimalnom vrijednosti $logW$ (slučaj kada je jedne klase ukupno $N/2$). Faktor normalizacije kod parametra maksimalno mogućeg stvarnog doprinosa iznad nasumične točnosti ($\Delta Q_{2,max}$), u pravilu je 0.5, što odgovara najsloženijem mogućem modelu. Normalizirani parametri relativna su mjera složenosti varijable na temelju koje su definirani pragovi (razine) složenosti. Ako je složenost neke varijable ispod odabranog praga, varijabla bi trebala biti isključena iz modela i razmatranja u postupku modeliranja. Za takvu varijablu možemo reći i da nije dovoljno informativna, tj. da sadrži malu količinu korisne informacije.

Odabrani normirani parametri složenosti prikazani su u obliku mrežnog poslužitelja za besplatno korištenje u analizi složenosti varijabli (*engl.* Classification variable complexity parameter estimator). Aplikacija se može koristiti i na izračun složenosti klasifikacijskih varijabli s dva stanja, ali i na varijablama s kontinuiranim vrijednostima koje automatski dihotomizira. Razvijena aplikacija dostupna je online na adresi http://meteo2.irb.hr/shiny/CA/.

Primjeri korisne uporabe razvijenih parametara i rezultata dobivenih u disertaciji prikazani su na više skupova podataka u QSAR modeliranju topljivosti i biološke aktivnosti organskih kemijskih spojeva, te u primjeni u rangiranju modela u predikcijskim natjecanjima različitih grupa u modeliranju u bioinformatici. Primjene rezultata istraživanja na kontinuirane varijable omogućena je dihotomizacijom (digitalizacijom) varijable koja se provodi u odnosu na srednju vrijednost svih njenih vrijednosti. Nakon toga, primjenjuje se identičan postupak kao i za binarne klasifikacijske varijable. Usporedbom modificirane normalizirane Boltzmannove formule za entropiju ($logW_{norm}$) i normalizirane maksimalne stvarne točnosti ($\Delta Q_{2,max,norm}$) zaključeno kako je normalizirana entropija manje strog kriterij u procjeni složenosti varijable.

Istražena je korist primjene parametra u rangiranju klasifikacijskih modela za predviđanje ekspresije tumora na temelju informacija o utvrđenim genskim mutacijama. Pokazalo se da parametar ΔQ_2 favorizira one modele koji imaju slične vrijednosti dviju vrsta pogrešaka ($u \sim o$) u odnosu na druge parametre za iskazivanje kvalitete modela poput *MCC*, *F1* i Q_2. Osim toga parametar definiran je u svim slučajevima (za bilo koju vrijednost *p, n, u* i *o*), i nije osjetljiv na raspodjelu podataka, za razliku od parametara *MCC* i *F1*.

Izrazi izvedeni u disertaciji za karakteristične vrijednosti parametara za iskazivanje točnosti klasifikacijskih modela te za složenost klasifikacijskih varijabli dosad nisu objavljeni u znanstvenoj literaturi, i predstavljaju originalni doprinos. Također, simulacijama je potvrđeno i da su najvjerojatnija nasumična vrijednost bilo kojeg od spomenutih parametara zapravo srednja vrijednost svih vrijednosti tog parametra koji se dobije iscrpnim permutacijskim simulacijama (pri kojima je varijabla *E* u stalnom poretku, a varijabla *M* permutira se u svim poretcima).

U slučaju stvarnih modela, izmjenjive varijable su manje česte pa je u skladu s time napravljena i generalna tablica supstitucija za dobivanje karakterističnih vrijednosti parametara (*Prilog 3.36*) koja vrijedi za sve binarne varijable neovisno o udjelu klasa, a primjenjiva je i na izmjenjivim varijablama za određivanje karakterističnih vrijednosti funkcija parametara. Pomoću nje moguće je dobiti općenitije formule za karakteristične vrijednosti funkcija drugih parametara (*Prilog 3.37*).

6. LITERATURA

[1] C. Hansch, T. J. Fujita, p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure, J. Am. Chem. Soc., 1964, 86, 1616

[2] C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, F. Geiger, M. Streich, The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients, J. Am. Chem. Soc., 1963, 85, 2817-2824

[3] B. Anfinsen, E. Haber, M. Sela, F. H. White Jr., The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain, Proc. Natl. Acad. Sci., 1961, 47, 1309-1314

[4] I. Bošnjak, V. Bojović, T. Šegvić-Bubić, A. Bielen, Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank, Prot. Enginer. Des. & Sel., 2014, 27, 65-72

[5] S. M. Free Jr., J. W. Wilson, A Mathematical Contribution to Structure-Activity Studies, J. Med. Chem., 1964, 7, 395-399

[6] Talete srl, DRAGON 5.4, http://www.talete.mi.it, (accessed June 23, 2019)

[7] H. Van De Waterbeemd, R. E. Carter, G. Grassy, H. Kubiny, Y. C. Martin, M. S. Tute, P. Willett, Pure&Appl. Chem., Glossary of terms used in computational drug design, (IUPAC recommendations 1997), 1997, 69, 1137-1152

[8] D. Weininger, SMILES, A Chemical Language and Information System, 1. Introduction to Methodology and Encoding Rules, J. Chem. Inf. Comput. Sci., 1988, 28,

[9] H. Öztürk, Text-based Machine Learning Methodologies for Modelling drug-target Interactions, Submitted to the Institute for Graduate Studies in Science and Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Graduate Program in Computer Engineering, Bŏgaziçi University 2019

[10] J. Sadowski, J. Gasteiger, G. Klebe, J., Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures, J. Chem. Inf. Model., 1994, 34, 1000–1008

[11] Molecular Networks GmbH, Nuremberg, Germany, 3D Structure Generator CORINA Classic, www.mn-am.com, Accessed 19.3.2020

[12] Online SMILES Translator and Structure File Generator, https://cactus.nci.nih.gov/translate/, (accessed January 19, 2020)

[13] Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. , J. Chem. Inf. Model., 1992, 32, 244

[14] N. M. O'Boyle, M. Banck, C. A James, C. M., T. Vandermeersch, G. R Hutchison, Open Babel: An open chemical toolbox, J. Cheminformatics, 2011, 3

[15] Open Babel: Chemical file format converter, http://www.cheminfo.org/Chemistry/Cheminformatics/FormatConverter/index.html, (accessed March 19, 2020)

[16] OECD, Guidance Document on the Validation of [(Q)SAR] Models, ENV/JM/MONO, 2017, 2

[17] I. Gutman, N. Trinajstić, Graph theory and molecular orbitals. Total π-electron energy of alternant hydrocarbons, Chem. Phys. Lett., 1972, 17, 535-538

[18] M. Randić, On characterizing molecular branching, J. Am. Chem. Soc., 1975, 97, 6609-6615

[19] N. Trinajstić, Chemical Graph Theory, CRC: Boca Raton, FL, 1992, 2nd edition,

[20] J. J. P. Stewart, MOPAC (Molecular Orbital PACkage), Stewart Computational Chemistry,

Colorado Springs, CO, 2016
[21] M. J. Frisch, et al., Gaussian 09, Revision A.02, Gaussian, Inc.,Wallingford, CT, 2009
[22] L. Kuo, B. Mallick, Variable selection for regression models, Sankhyā: The Indian Journal of Statistics Series B (1960-2002), 1960–2002, 60, 65–81
[23] S. Garavaglia, A. Sharma, A Smart Guide to Dummy Variables: Four Applications and a Macro, https://stats.idre.ucla.edu/wp-content/uploads/2016/02/p046.pdf, (accessed June 5, 2020)
[24] R. Todeschini and V.Consonni, Molecular Descriptors for Chemoinformatics, WILEY-VCH, Weinheim (Germany), 2009, 2, 1257
[25] D. R. Rogers, A. J. Hopfinger, Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships, J. Chem. Inf. Comput. Sci., 1994, 34, 854-866
[26] B. Lučić, N. Trinajstić, Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling, J. Chem. Inf. Comput. Sci., 1999, 39, 121–132
[27] T. A. Andrea, H. Kalayeh, Applications of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors, J. Med. Chem., 1991, 34, 2824-2836.
[28] C. Pappin, P. Hojrupt, A. J. Bleasbyl, Rapid identification of proteins by peptide-mass fingerprinting, Curr. Biol., 1993, 327-332
[29] Daylight Chemical Information Systems, Inc, Daylight theory: Fingerprints, https://www.daylight.com/dayhtml/doc/theory/theory.finger.html, (accessed January 20, 2020)
[30] A. Tropsha, W. Zhang, Identification of the descriptor pharmacophores using variable selection QSAR: Application to database mining, Curr. Pharm. Des., 2001, 7, 599-612
[31] J. G. Topliss, R. J. Costello, Chance correlations in structure-activity studies using multiple regression analysis, J. Med. Chem. , 1972, 15, 1066-106
[32] J. G. Topliss, R. P. Edwards, Chance factors in studies of quantitative structure activity relationships, J. Med. Chem. , 1979, 22, 1238-124
[33] D. M. W. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation, J. Machine Learning Techn., 2011, 2, 37-6
[34] J. Batista, D. Vikić-Topić, B. Lučić, The Difference Between the Accuracy of Real and the Corresponding Random Model is a Useful Parameter for Validation of Two-State Classification Model Quality, Croat. Chem. Acta, 2016, 89
[35] B. Lučić, J. Batista, V. Bojović, M. Lovrić, A. Sović Kržić, D. Bešlo, D. Nadramija, D. Vikić-Topić, Estimation of random accuracy and its use in validation of predictive quality of classification models within predictive challenges , Croat. Chem. Acta, 2019, 92, 379–39
[36] J. S. Armstrong, F. Collopy, Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons, Int. J. Forecast, 1992, 8, 69–8
[37] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (*MCC*) over *F1* score and accuracy in binary classification evaluation, BMC Genomics, 2020, 21
[38] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit., Psychol. Bull., 1968, 70, 213–22
[39] P. Diaconis, D. A. Freedman, De Finetti's generalisations of exchangeability. In Jeffrey, R.C. (Ed.) Studies in inductive logic and probability (Vol. II), pp. 233–249, University of California Press, 1980
[40] Wikipedia contributors, Exchangeable random variables, Wikipedia - The Free Encyclopedia, https://en.wikipedia.org/wiki/Exchangeable_random_variables, (accessed August 21, 2020)

[41] S. Chasalow, combinatorics utilities, San Francisco (CA), Github, https://cran.r-project.org/package=combinat, (accessed August 11, 2020)
[42] L. Tierney, The R Compiler Package, https://stat.ethz.ch/R-manual/R-devel/library/compiler/html/00Index.html, (accessed August 11, 2020)
[43] M. Dowle, A. Srinivasan, J. Gorecki , M. Chirico, P. Stetsenko et. al., data.table: Extension of 'data.frame', San Francisco (CA), Github, https://cran.r-project.org/web/packages/data.table/
[44] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, https://ggplot2.tidyverse.org, (accessed April 15, 2020)
[45] G. Grothendieck, Manipulate R Data Frames Using SQL, San Francisco (CA), Github, https://cran.r-project.org/web/packages/sqldf/sqldf.pdf, (accessed August 8, 2020)
[46] U. Ligges, M. Maechler, S. Schnackenberg, scatterplot3d: 3D Scatter Plot, San Francisco (CA), Github, https://cran.r-project.org/web/packages/scatterplot3d/index.html, (accessed August 11, 2020)
[47] B. Hamner, M. Frasco, E. LeDell, Metrics: Evaluation Metrics for Machine Learning, San Francisco (CA), Github, https://cran.r-project.org/web/packages/Metrics/index.html, (accessed August 11, 2020)
[48] H. Bengtsson, R.utils: Various Programming Utilities, San Francisco (CA), Github, https://cran.r-project.org/web/packages/R.utils/index.html, (accessed August 11, 2020)
[49] W. Chang, R6: Encapsulated Classes with Reference Semantics, San Francisco (CA), Github, https://cran.r-project.org/web/packages/R6/index.html, (accessed August 11, 2020)
[50] G.J. Schutten, C.H. Chan, T. J. Leeper, J. Foster, et al, readODS: Read and Write ODS Files, San Francisco (CA), Github, https://cran.r-project.org/web/packages/readODS/index.html, (accessed August 11, 2020)
[51] J. Godden, J. Bajorath, An Information-Theoretic Approach to Descriptor Selection for Database Profiling and QSAR Modeling, QSAR Comb Sci, 2003, 22, 487-497.
[52] M. G. Sobell, A practical guide to Ubuntu Linux, Pearson Education, 2015
[53] V. Bojović, B. Lučić, D. Bešlo, K. Skala, N. Trinajstić, Calculation of Topological Molecular Descriptors Based on Degrees of Vertices, 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019 , 266–269; server dostupan na http://meteo2.irb.hr/indexer/
[54] K. Arnold, J. Gosling, D. Holmes, The Java programming language, Addison Wesley Professional, 2005
[55] Apache software foundation, Apache maven, http://maven.apache.org/, (accessed April 15, 2020)
[56] Bootstrap Core Team, CSS - Bootstrap, https://getbootstrap.com/docs/3.4/css/, (accessed April 15, 2020)
[57] The jQuery Team, Fast, small, and feature-rich JavaScript library, https://jquery.com/, (accessed January 4, 2020)
[58] R Core Team, The R Project for Statistical Computing, http://www.R-project.org/, (accessed January 4, 2020)
[59] K. Soetaert, Plotting Multi-Dimensional Data, San Francisco (CA), Github, https://cran.r-project.org/web/packages/plot3D/index.html, (accessed March 31, 2020)
[60] Origin(Pro), Version Number (e.g. "Version 2020"). OriginLab Corporation, Northampton, MA, USA.
[61] Microsoft Corporation, Microsoft excel, https://office.microsoft.com/excel

[62] RStudio, Inc, Shiny: Easy web applications in R, http://shiny.rstudio.com, (accessed April 15, 2020)
[63] D. Juretić, D. Vukičević, D. Petrov, M. Novković, V. Bojović, B. Lučić, N. Ilić & A. Tossi, Knowledge-based computational methods for identifying or designing novel, non-homologous antimicrobial peptides, Eur. Biophys. J., 2011, 40, 371-385
[64] RStudio Team, RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, http://www.rstudio.com/, (accessed April 15, 2020)
[65] M. Novković, J. Simunić, V. Bojović, A. Tossi, D. Juretić, DADP: the Database of Anuran Defense Peptides, Bioinformatics, 2012, 28, 1406–1407
[66] J. Huuskonen, Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology, J. Chem. Inf. Comput. Sci., 2000, 40, 773–777
[67] R.P. Verma, C. Hansch, QSAR modeling of taxane analogues against colon cancer, Eur. J. Med. Chem., 2010, 45, 1470–1477
[68] B. Rost,C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure., Proteins, 1994, , 55-72
[69] C. E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, 1948, 27, 379-423
[70] C. E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, 1948, 27, 623–656
[71] J. Dutka, The early history of the factorial function, Archive for History of Exact Sciences, 1991, 43, 225–249
[72] E. T. Jaynes, Gibbs vs Boltzmann entropies, Am. J. Phys, 1965, 33, 391-398
[73] B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta. - Prot. Struc., 1975, 405, 442–451
[74] Q. Zhu, On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset, Pattern Recognition Letters, 2020, 136, 71–80
[75] L. A. Jeni, J. F. Cohn, F De La Torre, Facing Imbalanced Data--Recommendations for the Use of Performance Metrics, 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, ,
[76] V. Bojović, Classification variable complexity parameter estimator, http://meteo2.irb.hr/shiny/CA/, (accessed March 12, 2020)
[77] R. Liu, S.-S. So, Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility, J. Chem. Inf. Comput. Sci., 2001, 41, 1633–1639
[78] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview , Bioinformatics, 2000, 16, 412–424
[79] SMC-DNA Challenge Participants, C. I. Cooper, D. Yao, D. H. Sendorek, T. N. Yamaguchi, C. P'ng, K. E. Houlahan, C. Caloian, M. Fraser, K. Ellrott, A. A. Margolin, R. G. Bristow, J. M. Stuart, P. C. Boutros, Valection: design optimization for validation and verification studies, BMC Bioinformatics, 2018, 19, 339
[80] ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, A. D. Ewing, K. E. Houlahan, Y. Hu, et al., Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection, Nat. Methods., 2015, 12, 623–630
[81] D. Chicco, Ten quick tips for machine learning in computational biology, BioData Min., 2017, 10, 35

[82] V. Bojović, ProtSeqAnalizer, http://meteo2.irb.hr/doktorat/ProtSeqAnalzyer.zip, (accessed July 19, 2020)
[83] W. Antweiler, Fun with indicator variables, https://wernerantweiler.ca/blog.php?item=2015-06-26
[84] R. Todeschini, V. Consonni, A. Maiocchi, The K correlation index: theory development and its applications in chemometrics. Chemom. Intell. Lab. Syst., 1998, 46, 13–29
[85] J.R. Cano, Analysis of data complexity measures for classification, Expert Systems with Applications, 2013, 40, 4820–4831
[86] B. Senliol, G. Gulgezen, L. Yu, Z. Cataltepe, Fast Correlation Based Filter (FCBF) with a Different Search Strategy, 2008 23rd International Symposium on Computer and Information Sciences, 2008, 43834
[87]-M. Hall, Correlation-based Feature Selection for Machine Learning (PhD thesis), University of Waikato, 1999
[88] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, et. al., A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches, Ieee Transactions On Systems, Man, And Cybernetics - part C: Applications And Reviews, 2012, 42, 463 - 484
[89] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, 2019, 91, 216–231
[90] The DREAM Consortium, Dream challenges, http://dreamchallenges.org, (accessed June 23, 2020)
[91] NCI DREAM Community, J. C. Costello, L. M. Heiser, E. Georgii, et al., A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms, Nat. Biotechnol., 2014, 32, 1202–1212
[92] GNU, P, Free Software Foundation. Bash (3.2. 48)[Unix shell program], 2007
[93] R. Bocinsky, pkgTest, Install And Load A Package, https://www.rdocumentation.org/packages/FedData/versions/1.1.0/topics/pkgTest, (accessed April 27, 2020)
[94] A. S. Kagel, counting sort, U.S. National Institute of Standards and Technology, Dictionary of Algorithms and Data Structures, 2006, https://xlinux.nist.gov/dads/HTML/countingsort.html, (accessed October 25, 2020)
[95] O. Mersmann, microbenchmark, Infrastructure to accurately measure and compare the execution time of R expressions, San Francisco (CA), Github, https://github.com/joshuaulrich/microbenchmark/, (accessed April 15, 2020)
[96] W. Chang, RStudio et.al, Shiny themes, https://cran.r-project.org/web/packages/shinythemes/index.html, (accessed August 22, 2020)
[97] R Studio Team, DT: An R interface to the DataTables library, https://rstudio.github.io/DT/, (accessed January 2, 2020)
[98] G. Csardi, Cross-Platform 'zip' Compression, https://www.rdocumentation.org/packages/zip/versions/2.1.1, (accessed August 22, 2020)

7. SAŽETAK

Kako bi se iz strukture dobila saznanja o svojstvima molekula, potrebno je tu strukturu opisati različitim strukturnim varijablama iz kojih se dobivaju različiti modeli odnosa strukture i svojstava. U takvim modelima važno pitanje je procjena razine nasumične korelacije (ili podudarnosti) koja je prisutna u svakom modelu. U slučaju ovoga rada nasumična korelacija odnosi se na vrijednost dobivenu koreliranjem varijable (*E*) s nasumičnom permutacijom same sebe (*M*), što je ekvivalentno matematičkom konceptu izmjenjivih varijabli *E* i *M*. Poželjno je da model ima mali broj značajnijih strukturnih varijabli koje sadrže što više korisne informacije o strukturi molekula. Stoga, pri izboru deskriptora u konačni model korisno bi bilo identificirati deskriptore koji imaju niski informacijski sadržaj, i izuzeti ih iz daljnjih analiza. To bi i ubrzalo postupke modeliranja jer bi se razmatrao manji početni skup strukturnih molekularnih deskriptora (varijabli).

U ovom istraživanju korištene su varijable s dvije klase, tj. klasa „0“ i klasa 1, a uspostavljena je i analogija s rezultatima za druge vrste varijabli. Analiza je napravljena za izmjenjive varijable (one varijable koje su nastale permutacijom postojeće varijable) i ona je rezultirala formulama za izmjenjive varijable koje se mogu poopćiti i primijeniti u izračunu podudarnosti općenitih parova binarnih varijabli. Dobivene su supstitucijske relacije koje vrijede za svaki parametar za izmjenjive varijable, i pomoću njih moguće je računati razne karakteristične funkcije parametara kao što su minimalna, maksimalna i prosječne nasumična vrijednosti, te njihovi rasponi.

Izvedene formule provjerene su simulacijama. Simulacije su provedene s varijablama veličine $N = 100$ i to s 100 000 parova izmjenjivih varijabli *E* i *M* za svaki udio klase 1 (x, %) kojem je vrijednost bila od 1 % do 99 % uz korak od 1 %. U prijašnjim radovima nasumične prosječne vrijednosti odnosile su se na najvjerojatnije nasumične vrijednosti, ali se pokazalo da to vrijedi samo u nekim slučajevima, dok prosječna vrijednost vrijedi općenito pa je došlo do korekcije naziva parametra.

Informacijski sadržaj u nekoj varijabli *E* u disertaciji povezan je se s njenom složenošću, tj. s brojem mogućih nasumičnih permutacija varijable, te s iznosom najmanje, najveće i prosječne nasumične podudarnosti koja se može dobiti usporedbama (svaki put) originalnog poretka varijable *E* s jednom od njenih neidentičnih permutacija *M*.

Složenost varijable određivana je pomoću modificirane Boltzmannove formule entropije i prilagodbom postojeće karakteristične vrijednosti ΔQ_2 parametra koji se računa pomoću udjela klase 1 (x) u varijabli. Provedene su analize karakterističnih vrijednosti parametara kao što su $Q_2, \Delta Q_2, MAE, s$, $MCC, F1$ i κ za koje su dobiveni izrazi za izračun minimalnih, maksimalnih i prosječnih nasumičnih vrijednosti te njihovih raspona. Korelacije između karakterističnih vrijednosti parametara te njihovih raspona s entropijom varijable omogućile su odabir najboljeg kandidata kao dodatnu mjeru složenosti, te je tako je pored entropije *logW* za procjenu i analizu složenosti odabrana maksimalna vrijednost parametara ΔQ_2 ($\Delta Q_{2,max}$).

Dobiveni rezultati i zaključci korišteni su za definiranje postupka za provjeru kvalitete modela na primjerima skupova molekula iz koji su priređeni radom na disertaciji, a i na skupovima iz literature. Naposljetku, razvijen je mrežni poslužitelj za potrebe izračuna vrijednosti složenosti i nivoa složenosti, te nasumične korelacije klasifikacijskih varijabli. Osim određivanja karakterističnih vrijednosti parametara i složenosti klasifikacijskih varijabli s dva stanja, poslužitelj

se može koristiti i za rad s kontinuiranim varijablama nakon provedbe automatske dihotomizacije u odnosu na srednju vrijednost varijable..

Također, razvijen mrežni poslužitelj za analizu složenosti varijabli provjeren je na skupovima deskriptora organskih spojeva i proteina izračunanih s pomoću programa razvijenih u disertaciji. Za tu svrhu razvijen je mrežni poslužitelj za izračun različitih topoloških deskriptora, te aplikacija za izračun različitih strukturnih motiva iz primarne strukture skupa proteina.

8. SUMMARY

In order to gain knowledge about the properties of molecules from a structure, it is necessary to describe the structure with different structural variables from which different models of the relationship between structure and properties are obtained.

In such models, an important issue is the assessment of the level of chance correlation present in each model. In the case of this paper, chance correlation refers to the correlation value obtained by correlating a variable with a random permutation of itself. Preferably, the model has to have a small number of significant structural variables that contain as much useful information as possible about the structure of the molecules. Therefore, when selecting descriptors in the final model, it would be useful to identify descriptors that have low information content and to exclude them from further analyzes. That would also speed up modeling procedures because a smaller initial set of structural molecular descriptors (variables) would be considered. In this research, variables with two classes were used, i.e. class "0" and class "1", and an analogy with the results for other types of variables was established. The analysis was made for exchangeable variables (those variables that were created by permutation of an existing variable) and resulted in formulas for both exchangeable variables and all binary variables in general. Substitutions valid for each parameter of variable variables were obtained, such as minimum, maximum, range, and random mean functions, as well as their ranges.

Derived formulas are verified by simulations. Simulations were done by variables of size $N = 100$ which produced 100,000 pairs of variables *E* and *M* for each content of class 1 (x) whose value ranged from 1 % to 99 % with a step of 1 %. In recent papers average random values were associated by the most probable random values, but that assumption works only in specific cases, while average random value is a valid term in general,. The information contained in a variable *E* in the dissertation is related to its complexity, *i.e.* with the number of possible random realizations of the variable, as well as with the minimal, maximal, and the average random agreement (correlation) that can be obtained with a variable by comparing (each time) of original order of values of variable *E* by one of it's nonidentical permutations *M*.

The complexity of the variable is determined using the modified Boltzmann entropy formula and by the adjusted characteristic value ΔQ_2 of the accuracy parameter calculated using the content of class 1 (x) in the variable. In addition, the analyzis of characteristic values of other parameters such as s, Q_2, ΔQ_2,$Q_{2,rnd}$, MCC, $F1$ were made, for which the formulas for minimum, maximum, range, and random mean functions values are obtained.

The correlations between the characteristic values of parameters and their ranges with the entropy of variable enabled the selection of the best candidate as an aditional measure of complexity. Thus, due to high correlation with the entropy ($logW$), the maximum value of the parameter ΔQ_2 ($\Delta Q_{2,max}$), which depends on the content of class "1" (x), was selected as an additional complexity measure.

The obtained results and conclusions were used to define the procedure for checking the quality of models using the examples of sets of molecules from which they were prepared by working on the dissertation, as well as on sets from the literature. Finally, a network server was developed for the purpose of calculating complexity values and complexity levels, and random correlations of classification variables. In addition to determining the characteristic values of parameters and the

complexity of binary variables, the server has the ability to work with continuous variables after their automatic dichotomization relative to their mean values. Also, the developed server for analysis of complexity of variables was tested on data sets of descriptors of organic compounds and proteins calculated by the computer programs developed in the dissertation. For that purpopse a network server for calculation of various topological descriptors for a set of compounds was developed, as well as the application for extraction of structural motives from primary structure of a set of proteins.

9. POPIS KRATICA

Kratica	Značenje
A	Oznaka apsolutne vrijednosti u indeksu parametra
abs	Apsolutna vrijednost
AA	Oznaka u indeksu parametra koja se odnosi na uzlazni poredak varijabli E (eksperimentalna) i M (modelna)
AD	Oznaka u indeksu parametra koja se odnosi na suprotan poredak varijabli E (eksperimentalna) i M (modelna)
CAS	Serijski broj molekule (identifikator) koji određuje „Chemical Abstracts Service" dio organizacije „American Chemical Society"
CIF	Crystallographic Information File
E	Oznaka eksperimentalne varijable
$entX$	Binarna entropija
FASTA	Zapis datoteke koji sadrži sekvence proteina, DNA ili RNA
F_1score	F_1score parametar (mjera točnosti testa)
$F1$	F_1score parametar
L	Oznaka u indeksu parametra za slučaj kada je $x \leq \frac{1}{2}$
$logW$	Entropija (logaritam broja kombinacija)
M	Oznaka modelirane varijable
MAE	Prosječna pogreška srednje vrijednosti
MCC	Matthewsov koeficijent korelacije
Max	Oznaka maksimalne vrijednosti varijable
Min	Oznaka minimalne vrijednosti varijable
MOL	Zapis datoteke koji sadrži informacije o koordinatama atoma i vezama
MOL/SDF	Kombinacija MOL i SDF zapisa datoteke
n	Ukupni broj točno predviđenih podataka klase 0 (TN – engl. *true negative prediction*)
N	Veličina varijable
o	Ukupni broj netočno predviđenih podataka klase 0 (FP – engl. *False positive prediction*)
OOP	Objektno orjentirana paradigma – način organiziranja programskog koda
p	Ukupni broj točno predviđenih podataka klase 1 (TP – engl. *true positive prediction*)
PDB	Protein data bank – vrsta datotetke za pohranu eksperimentalnih podataka o svakom atomu makromolekule
Q_2	Točnost modela
$Q_{2,rnd}$	Najvjerojatnija nasumična točnost
QSAR	Quality structure activity relationship
QSPR	Quality structure property relationship
R	Oznaka u indeksu parametra za slučaj kada je $x \geq \frac{1}{2}$
R	Programski jezik namijenjen za statističke izračune

r	Pearsonov koeficijent korelacije
rnd	Oznaka srednje nasumične vrijednosti parametra (u indeksu)
$RMSE$	Root means square error parametar
s	Standardna pogreška parametar
S	Simulacija (oznaka u indeksu)
SDF	Structure data file – vrsta datoteke za pohranu dodatnih informacija o molekulama
SE	Standardna pogreška srednje vrijednosti
sim	Simulacijska vrijednost
SMILES	„Simplified molecular-input line-entry system" - format zapisa strukture molekula
SQL	Structured query language – programski jezik za rad s bazama podataka
u	Ukupni broj netočno predviđenih podataka klase 1 (FN – engl. *False negative prediction*)
x	Udio klase 1 u N
xN	Količina klase 1 u N za slučaj izmjenjivih varijabli, dok kod varijabli općenito je količina klase 1 u varijabli E
X	Količina klase 1 u N
yN	Količina klase 1 u M varijabli kod asimetričnih varijabli
y	Oznaka udjela klase 1 u varijabli M kod varijabli gdje nisu nužno omjeri klasa jednaki
Δ	Oznaka raspona
$\Delta(\Delta Q_2)$	Raspon $\Delta(Q_2)$ parametra
$\Delta(F_1 score)$	Raspon $F_1 score$ varijable
$\Delta(\kappa)$	Raspon Cohenovog kapa parametra
$\Delta(MAE)$	Raspon MAE parametra
$\Delta(MCC)$	Raspon MCC parametra
ΔQ_2	Doprinos modela
$\Delta(Q_2)$	Raspon Q_2 parametra
$\Delta(Q_{2,rnd})$	Raspon $Q_{2,rnd}$ parametra
$\Delta(s)$	Raspon s parametra
κ	Cohenov kapa parametar
κ_{rnd}	Prosječna nasumična vrijednost Cohenovog kapa parametra
σ	Standardna devijacija

10. PRILOZI

PRILOG 1 (1. Uvod)

PRILOG 1.1. Zapis strukture dietil-fenil-fosfata sa *Slike 1.1* u obliku MOL/SDF.*****

```
OpenBabel01182010332D

 15 15  0  0  0  0  0  0  0  0999 V2000
   -1.0000    0.7321    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5000   -0.1340    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0000   -1.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -1.0000    0.0000 P   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -2.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.0000   -1.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000   -1.8660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5000   -1.8660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.8660    0.5000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.8660    1.5000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7321    2.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981    1.5000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981    0.5000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7321    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  4  5  2  0  0  0  0
  4  6  1  0  0  0  0
  4  9  1  0  0  0  0
  6  7  1  0  0  0  0
  7  8  1  0  0  0  0
  9 10  1  0  0  0  0
 10 15  1  0  0  0  0
 10 11  2  0  0  0  0
 11 12  1  0  0  0  0
```

```
 12 13  2  0  0  0  0
 13 14  1  0  0  0  0
 14 15  2  0  0  0  0
M  END
$$$$
```

***** SDF datoteka sadrži više struktura zapisanih u MOL obliku, a međusobno su odvojene linijom u kojoj je znak „$$$$“.

U MOL/SDF zapisu u prvoj liniji nalazi se informacija o programu kojim je datoteka dobivena, nakon toga slijede podaci o x, y, i z koordinatama atoma u molekuli iskazanu u Ångstromima nakon kojih slijedi stupac s oznakama vrste atoma, U slučaju da informacije o koordinatama nisu dostupne, u pripadajućim stupcima bit će nule. Ukoliko je riječ o 2D strukturama, nule će biti upisane u trećem stupcu koji odgovara koordinati z. Potom slijede linije s informacijama o vezama među atomima (povezanost) gdje dva broja predstavljaju redni broj atoma (redom kako su navedeni iznad u linijama s oznakama atoma) koji su međusobno povezani. Treći stupac sadrži informaciju o vrsti kovalentne veze, tj. je li veza jednostruka (broj 1), dvostruka (broj 2) itd.

Datoteka završava sa znakom $$$$.

PRILOG 2 (2. Metode)

PRILOG 2.1 Dokaz pravila $o = u$ za simulacijske varijable

$$X_E = \sum_{i=1}^{N} E_i \,, \forall E_i = 1$$

$$X_M = \sum_{i=1}^{N} M_i \,, \forall M_i = 1$$

$$X = xN$$

$$x_E = \frac{p + u}{N}$$

$$x_M = \frac{p + o}{N}$$

$$x_E = x_M$$

$$p + u = p + o$$

$$o = u$$

***** X je broj jedinica u eksperimentalnoj (E) ili modeliranoj (M) varijabli, dok je $x$oznaka udjela klase 1 u varijabli E ili M

Ovim se izvodom dokazuje da kod izmjenjivih varijabli vrijedi pravilo $o = u$. Ovdje su u i o elementi tablice kontigencije, gdje u predstavlja ukupan broj slučajeva kad je klasa 1 predviđena kao klasa 0, dok o predstavlja ukupni broj slučajeva kad je klasa 0 predviđena kao klasa 1).

X predstavlja ukupan broj slučajeva kad varijabla E (eksperimentalna) ili M(modeliranoj) poprimaju vrijednost jednaku 1 (klasa 1). a x je udio klase 1 u varijabli duljine $N, x = X \;\; N$.

Kod izmjenjivih varijabli jednaki su udjeli klase 1 u varijabli E ($(p + u) \;\; N$)) i klase 1 u varijabli M $(p + o) \;\; N$. Stoga, izjednačavanjem imamo: $(p + u) \;\; N = (p + o)N$ pri čemu je p broj točno predviđenih vrijednost klase 1 u varijabli E s pomoću varijable M. Ili, drugim riječima, p je ukupni broj slučajeva kad je i u varijabli E i u varijabli M na istom mjestu vrijednost jednaka 1 (klasa 1). Iz gornje jednakosti proizlazi da je $o = u$. Ispod su dane definicije oznaka koje će se koristiti u disertaciji, te postupan matematički izvod ove jednakosti.

Tako za izmjenjive varijable vrijedi $N = p + o + u + n = \{o = u\} = p + 2u + n.$

U donjim izrazima p, n, u, o predstavljaju elemente matrice kontingencije. Kratice za parametre (mjere) kvalitete su: MCC predstavlja Matthewsov koeficijent korelacije, $Q_{2,rnd}$ prosječnu nasumičnu točnost, s standardnu pogrešku , $F1$ mjeru točnosti modela, MAE prosječnu apsolutnu pogrešku, ΔQ_2 doprinos modela i Q_2 točnost modela.

PRILOG 2.2 Pojednostavljenje formula koristeći pravilo $o = u$

$$MCC = \frac{np - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} = \frac{np - u^2}{\sqrt{(p+u)^2(n+u)^2}} = \frac{np - u^2}{(p+u)(n+u)}$$

$$Q_{2,rnd} = \frac{\frac{p+u}{N}\frac{p+o}{N} + \frac{n+o}{N}\frac{n+u}{N}}{N^2}(\%) = \frac{\left(\frac{p+u}{N}\right)^2 + \left(\frac{n+u}{N}\right)^2}{N^2}\ \%\quad [33]$$

$$s = \sqrt{\frac{o+u}{N}} = \sqrt{2\frac{u}{N}}$$

$$F_1 score = \frac{2p}{2p + o + u} = \frac{2p}{2p + 2u} = \frac{p}{p+u}$$

$$MAE = \frac{o+u}{p+u+n+o} = \frac{2u}{N}$$

$$\Delta Q_2 = Q_2 - Q_{2,rnd}(\%) = \frac{p+n}{N} - \frac{(p+u)^2 + (n+u)^2}{N^2} = N\frac{p+n}{N^2} - \frac{(p+u)^2 + (n+u}{N^2}$$

$$\Delta Q_2 = \frac{N(p+n) - (p+u)^2 - (n+u)^2}{N^2}$$

$$\kappa = \frac{\frac{p+n}{N} - \frac{\left(\frac{p+u}{N}\right)^2 + \left(\frac{u+n}{N}\right)^2}{N^2}}{1 - \frac{\left(\frac{p+u}{N}\right)^2 + \left(\frac{u+n}{N}\right)^2}{N^2}} = \begin{matrix} o = u \\ N = 2u + p + n \end{matrix} = \frac{np - u^2}{n+u\ \ p+u} = MCC$$

***** MCC predstavlja Matthewsov koeficijent korelacije, $Q_{2,rnd}$ prosječnu nasumičnu točnost, s standardnu pogrešku, F_1 mjeru točnosti modela, MAE prosječnu apsolutnu pogrešku, ΔQ_2 doprinos modela i Q_2 točnost modela. p – broj točno predviđenih podataka klase 1, n – broj točno predviđenih podataka klase 0, o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1

Primjenom pravila $o = u$ što vrijedi za izmjenjive varijable (ukupan broj netočnih predviđanja je jednak u obje varijable) pojednostavljuje formule.

PRILOG 2.3 Provjera formula za broj klasa

$$xN = p + u = p + o = X$$
$$n + u = N - xN = N(1 - x)$$

---

* xN – broj elemenata klase 1, X – broj elemenata klase 1, $N(1-x)$ – broj elemenata klase 0, N – ukupni broj podataka u varijabli, p – broj točno predviđenih podataka klase 1, n – broj točno predviđenih podataka klase 0, o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1

Ovim izvodom dokazano je da je ukupan broj klase 1 (xN) jednak duljini varijable umanjenoj za ukupan broj klase 0, tj. da je ukupan broj klase 0 $N(1-x)$.

PRILOG 2.4 Provjera pravila $p+u=xN$

$$p+u=\begin{matrix}0+xN=xN,\forall x\in[0,\frac{1}{2}]\\(2x-1)N+N(1-x),\forall x\in[\frac{1}{2},1]\end{matrix}=\begin{matrix}0+xN=xN,\forall x\in[0,\frac{1}{2}]\\2xN-N+N-xN,\forall x\in[\frac{1}{2},1]\end{matrix}$$

$$p+u=\begin{matrix}xN,\forall x\in[0,\frac{1}{2}]\\xN,\forall x\in[\frac{1}{2},1]\end{matrix}=xN,\forall x\in[0,1]$$

* xN – broj elemenata klase 1, X – broj elemenata klase 1, $N(1-x)$ – broj elemenata klase 0, N – ukupni broj podataka u varijabli. p – broj točno predviđenih podataka klase 1, n – broj točno predviđenih podataka klase 0, o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1

U izvodu je dokazano za poredak podataka „*AD*" (varijabla E uzlazno, varijabla M silazno) ne utječe na pravilo $p+u=xN$. Također, je dokazano da pravilo jednako vrijedi neovisno za slučajeve kad je udio $x\leq 1\ \ 2$ i kad je $x\ \geq\ 1\ \ 2$.

PRILOG 2.5 Provjera pravila $n+o=N(1-x)$

$$n+o=\begin{matrix}(1-2x)N+xN,\forall x\in[0,\frac{1}{2}]\\0+N(1-x),\forall x\in[\frac{1}{2},1]\end{matrix}=\begin{matrix}N-2xN+xN,\forall x\in[0,\frac{1}{2}]\\N(1-x),\forall x\in[\frac{1}{2},1]\end{matrix}$$

$$n+o=\begin{matrix}N-xN,\forall x\in[0,\frac{1}{2}]\\N(1-x),\forall x\in[\frac{1}{2},1]\end{matrix}=\begin{matrix}N(1-x),\forall x\in[0,\frac{1}{2}]\\N(1-x),\forall x\in[\frac{1}{2},1]\end{matrix}$$

$$n+o=N(1-x),\forall x\in[0,1]$$

* xN – broj elemenata klase 1, X – broj elemenata klase 1, $N(1-x)$ – broj elemenata klase 0, N – ukupni broj podataka u varijabli. p – broj točno predviđenih podataka klase 1, n – broj točno predviđenih podataka klase 0, o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1

Na isti način kao u *Prilogu 2.4*, dokazuje se da formula koja predstavlja udio klase 0 $n+o=N\ \ 1-x\ \ $ ostaje nepromijenjena bez obzira je li udio klase 1 u $Nx\ \geq\ 1\ \ 2$ ili $x\leq 1\ \ 2$.

PRILOG 3 (3. Rezultati)

Za izvođenje karakterističnih vrijednosti parametara Q_2, MAE, s, $MCC, F1$ i κ korištene su vrijednosti iz tablice supstitucijskih izraza (Materijali i metode - *Tablica 2.4*) za veličine p, n, u i o iskazane preko udjela klase 1 (x).

Parametar Q_2

PRILOG 3.1 Izvod izraza za minimalnu vrijednost parametra Q_2 (točnost/podudarnost permutirane varijable M u odnosu na varijablu E) za lijevi podinterval udjela klase 1, tj. za $x \leq 1\ \ 2$:

$$Q_{2,AD,L} = \frac{n+p}{N} = \frac{(1-2x)N+0}{N} = 1-2x, \forall x \in [0, \frac{1}{2}]$$

***** x – udio klase 1; N – ukupni broj podataka u varijabli; p – broj točno predviđenih vrijednosti klase 1; n – broj točno predviđenih vrijednosti klase 0; Q_2 je oznaka parametra točnosti/podudarnosti varijabli E i M; AD (*engl*. Ascending-Descending) - oznaka kada je E varijabla poredana nasuprotno u odnosu na varijablu M (varijabla E sortirana uzlazno, varijabla M sortiranu silazno); $L-$ oznaka za lijevi interval vrijednosti udjela klase 1, tj. za $x \leq 1\ \ 2$.

PRILOG 3.2 Izvod izraza za minimalnu vrijednost parametra Q_2 (točnosti/podudarnosti varijabli E i M) za desni podinterval udjela klase 1, tj. za $x\ \geq\ 1\ \ 2$

$$Q_{2,AD,R} = \frac{n+p}{N} = \frac{0+(2x-1)N}{N} = 2x-1, \forall x \in [\frac{1}{2}, 1]$$

***** značenja kratica x, N, p, n, Q_2 i AD objašnjena su ispod izvoda u *Prilogu 3.1*; $R-$ je oznaka za desni (*engl*. Right) interval vrijednosti udjela klase 1, tj. kada je $x\ \geq\ 1\ \ 2$.

PRILOG 3.3 Izvod izraza za maksimalnu vrijednost parametra Q_2 (točnosti/podudarnosti varijabli E i M) za cijeli interval udjela klase 1, tj. za $x \in [0,1]$

$$Q_{2,AA} = \frac{n+p}{N} = \frac{N(1-x)+xN}{N} = 1, \forall x \in [0,1]$$

***** značenja kratica x, N, p, n, i Q_2 objašnjena su ispod izvoda u *Prilogu 3.1*; AA (*engl*. Ascending-Ascending) - oznaka za slučaj kada su varijable E i M poredane (sortirane) uzlazno.

PRILOG 3.4 Izvod izraza za raspon Q_2 (parametar točnosti/podudarnosti varijabli E i M) za $x \leq \frac{1}{2}$

$$\Delta(Q_{2,L}) = Q_{2,AD,L} - Q_{2,AA} = 1-2x-1 = 2x, \forall x \in [0, \frac{1}{2}]$$

***** značenja kratica x, N, Q_2, AD i L objašnjena su ispod izvoda u *Prilogu 3.1*, a značenje kratice AA objašnjeno je ispod izvoda u *Prilogu 3.3*.

PRILOG 3.5 Izvod izraza za raspon Q_2 (parametar točnosti/podudarnosti varijabli E i M) za $x\ \geq\ 1\ \ 2$

$$\Delta(Q_{2,R}) = Q_{2,AD,R} - Q_{2,AA} = 1 - (2x - 1) = 2 - 2x = 2(1 - x), \forall x \in [\frac{1}{2}, 1]$$

***** značenja kratica x, N, Q_2, i AD objašnjena su ispod izvoda u *Prilogu 3.1*, značenja kratice R i AA objašnjena su ispod izvoda u *Prilogu 3.2* odnosno *Priloga 3.3.*

Parametar $Q_{2,rnd}$

PRILOG 3.6 Izvod izraza za minimum parametra $Q_{2,rnd}$ za izmjenjive varijable ($o = u$) za $x \leq 1\ \ 2$

Prosječna nasumična točnost modela $Q_{2,rnd}$ pojednostavljuje se supstitucijom $o = u$ u oblik dan u *Prilogu 3.6.*

$$Q_{2,rnd,AD,L} = \frac{p + u\ ^2 + \ \ n + u\ ^2}{N^2} = \frac{o + xN\ ^2 + \ \ \ 1 - 2x\ \ N + xN\ \ ^2}{N^2}$$

$$Q_{2,rnd,AD,L} = 2x^2 - 2x + 1, \forall x \in [0, \frac{1}{2}]$$

***** značenja kratica x, N, p, i n, objašnjena su ispod izvoda u *Prilogu 3.1*; $Q_{2,rnd}$- prosječna nasumična točnost; o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1; L je oznaka lijevog podintervala vrijednosti udjela klase 1 (x), tj. kada je $x \leq 1\ \ 2$; R je oznaka desnog podintervala vrijednosti udjela klase 1 (x), tj. kada je $x\ \geq\ 1\ \ 2$

Indeksom AD označen je uzlazni poredak za varijablu E, a silazni za varijablu M i služi za dobivanje minimuma bilo kojeg parametra. Poretkom AA varijabli E i M (obje varijable poredane u istom poretku) dobiva se <u>maksimalna</u> vrijednost parametara koji mjere točnost ili korelaciju: $Q_2, Q_{2,rnd}, \Delta Q_2, MCC, F1$ i κ. S druge strane, u tom slučaju dobiva se <u>minimalna</u> vrijednost parametara koji su mjere pogreške: MAE i s.

PRILOG 3.7 Izvod izraza za minimalnu vrijednost parametra $Q_{2,rnd}$ za izmjenjive varijable ($o = u$) za $x\ \geq\ 1\ \ 2$

$$Q_{2,rnd,AD,R} = \frac{(p + u)^2 + (n + u)^2}{N^2} = \frac{((2x - 1)N + N - xN)^2 + (0 + N - xN)^2}{N^2}$$

$$Q_{2,rnd,AD,R} = 2x^2 - 2x + 1, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*; R – oznaka desnog podintervala vrijednosti udjela klase 1 (x), tj. kada je $x\ \geq\ 1\ \ 2$

PRILOG 3.8 Izvod izraza za maksimalnu vrijednost parametra $Q_{2,rnd}$ za izmjenjive varijable $x \in [0,1]$

$$Q_{2,rnd,AA} = \frac{(p+u)^2 + (n+u)^2}{N^2} = \frac{(xN+0)^2 + (N(1-x)+0)^2}{N^2} = 2x^2 - 2x + 1, \forall x \in [0,1]$$

Usporedba konačnih izraza u *Prilozima 3.6 – 3.8* pokazuje da su minimalne vrijednosti u oba podintervala x i izaz za maksimalnu vrijednost u cijelom intervalu identični. Iz toga se može zaključiti kako je postupak izvođenja minimalnih vrijednosti za dva podintervala ispravan. Nadalje, pokazano je i da parametar prosječne nasumične točnosti ne ovisi o permutacijama, nego samo o udjelu klase 1 (x), jedne od dviju klasa (u ovom slučaju, kao referentna klasa izabrana je klasa 1).

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6.*

Parametar ΔQ_2

PRILOG 3.9 Izvod izraza za minimalnu vrijednost parametra ΔQ_2 za slučaj kada je $x \leq 1\ \ 2$

$$\Delta Q_{2,AD,L} = \frac{n+p}{N} - \frac{(p+u)^2 + (n+u)^2}{N^2} = 1 - \frac{(N(2x-1) - xN)^2 + (xN)^2}{N^2} - 2x = -2x^2$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*; ΔQ_2- doprinos modela, tj. iznos nađene podudarnosti (Q_2) varijabli E i M koji je iznad razine prosječne nasumične točnosti.

PRILOG 3.10 Izvod izraza za minimalnu vrijednost parametra ΔQ_2 za $x \ \geq\ 1\ \ 2$

$$\Delta Q_{2,AD,R} = \frac{n+p}{N} - \frac{(p+u)^2 + (n+u)^2}{N^2}$$

$$\Delta Q_{2,AD,R} = 1 - \frac{(N + N(2x-1) - xN)^2 + (N - xN)^2}{N^2} - 1$$

$$\Delta Q_{2,AD,R} = -2x^2 + 4x - 2 = -x(x-1)^2, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilozima 3.6* i *3.9*; R je oznaka desnog pod-intervala vrijednosti udjela klase 1 (x), tj. kada je $x\ \geq\ 1\ \ 2$.

PRILOG 3.11 Izvod izraza za maksimalnu vrijednost parametra ΔQ_2

$$\Delta Q_{2,AA} = \frac{n+p}{N} - \frac{(p+u)^2 + (n+u)^2}{N^2} = \frac{-N^2(x-1)^2 + (xN)^2}{N^2} - \frac{N(x-1) - xN}{N}$$

$$\Delta Q_{2,AA} = -2x(x-1)$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.9.*

PRILOG 3.12 Izvod izraza za raspon parametra ΔQ_2

$$\Delta(\Delta Q_2) = (Q_2 - Q_{2,rnd})_{AA} - (Q_2 - Q_{2,rnd})_{AD} = Q_{2,AA} - Q_{2,AD} = \Delta Q_2$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.9.*

Parametar *MAE*

PRILOG 3.13 Izvod izraza za maksimalnu vrijednost parametra MAE za $x \leq 1\ \ 2$

$$MAE_{AD,L} = \frac{o+u}{N} = \frac{2u}{N} = \frac{2xN}{N} = 2x, \forall x \in [0, \frac{1}{2}]$$

***** x – značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*; MAE– prosječna apsolutna pogreška

PRILOG 3.14 Izvod izraza za maksimalnu vrijednost parametra MAE za $x \geq 1\ \ 2$

$$MAE_{AD,R} = \frac{2u}{N} = \frac{2N - 2xN}{N} = 2 - 2x = 2(1-x), \forall x \in [\frac{1}{2}, 1]$$

*****značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.9.*; MAE– prosječna apsolutna pogreška

PRILOG 3.15 Izvod izraza za minimalnu vrijednost parametra MAE

$$MAE_{AA} = \frac{2u}{N} = \frac{2x0}{N} = 0, \forall x \in [0,1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.14.*

PRILOG 3.16 Izvod izraza za raspon parametra MAE za $x \leq 1\ \ 2$

$$\Delta(MAE_L) = MAE_{AD,L} - MAE_{AA} = 2x, \forall x \in [0, \frac{1}{2}]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.14.*

PRILOG 3.17 Izvod izraza za raspon parametra MAE za $x \geq 1\ \ 2$

$$\Delta(MAE_R) = MAE_{AD,R} - MAE_{AA} = 2 - 2x, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.14.*

Parametar *s*

PRILOG 3.18 Pojednostavljenje formule za parametar s (standardna pogreška – srednje kvadratno odstupanje) za izmjenjive varijable ($o = u$)

$$s = \overline{\frac{o+u}{N}} = \overline{2\frac{u}{N}}$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*; s – standardna pogreška

PRILOG 3.19 Izvod izraza za minimalnu vrijednost parametra *s*

$$s_{AA} = \overline{2\frac{u}{N}} = \overline{\frac{0}{N}} = 0, \forall x \in [0,1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.18.*

PRILOG 3.20 Izvod izraza za maksimalnu vrijednost parametra s za $x \leq 1\ \ 2$

$$s_{AD,L} = \overline{2\frac{u}{N}} = \overline{2\frac{xN}{N}} = \overline{2x}, \forall x \in [0,\frac{1}{2}]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.18.*

PRILOG 3.21 Izvod izraza za maksimalnu vrijednost parametra s za $x \geq 1\ \ 2$

$$s_{AD,R} = \overline{2\frac{u}{N}} = \overline{2\frac{N-xN}{N}} = \overline{2(1-x)}, \forall x \in [\frac{1}{2},1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.18.*

PRILOG 3.22 Izvod izraza za raspon vrijednosti parametra s za $x \leq 1\ \ 2$

$$\Delta(s_L) = s_{AD,R} - s_{AA} = \overline{2x} - 0 = \overline{2x}, \forall x \in [0,\frac{1}{2}]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.18.*

PRILOG 3.23 Izvod izraza za raspon vrijednosti parametra s za $x \geq 1\ \ 2$

$$\Delta(S_R) = S_{AD,R} - S_{AA} = \overline{2(1-x)} - 0 = \overline{2(1-x)}, \forall x \in [\frac{1}{2},1]$$

***** x – udio klase 1, N – ukupni broj podataka u varijabli, AD- oznaka kada je E varijabla poredana nasuprotno u odnosu na sortiranu M varijablu, R – oznaka za slučaj kada je $x \geq 1\ \ 2$, AA- oznaka za slučaj kada su E i M varijabla preslagane jednakim redoslijedom

Parametar *MCC*

PRILOG 3.24 Izvod MCC parametra za izmjenjive varijable ($o = u$)

$$MCC = \frac{np - ou}{\sqrt{(n+o)(n+u)(p+o)(p+u)}} = \frac{np - ou}{\sqrt{(n+o)^2(p+o)^2}} = \frac{np - ou}{(n+o)(p+o)}$$

$$MCC = \frac{np - o^2}{(n+o)(p+o)} = \{o = u\} = \frac{np - u^2}{(n+u)(p+u)}$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*; MAE– prosječna apsolutna pogreška, MCC- Matthewsov koeficijent korelacije

PRILOG 3.25 Izvod izraza za maksimalnu vrijednost parametra MCC za $x \geq 1\ \ 2$

$$MCC_{AA} = MCC_{max} = \frac{np - u^2}{(1-x)xN^2} = \frac{N(1-x)xN - 0}{(1-x)xN^2} = 1, \forall x \in [0,1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.24*

PRILOG 3.26 Izvod izraza za minimalnu vrijednost parametra MCC za $x \leq 1\ \ 2$

$$MCC_{AD,L} = MCC_{min} = \frac{np - u^2}{(1-x)xN^2} = \begin{matrix} p = 0 \\ n = (1-2x)N \\ u = xN \end{matrix} = \frac{x}{x-1}, \forall x \in [0, \frac{1}{2}]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.24*

PRILOG 3.27 Izvod izraza za minimalnu vrijednost parametra MCC za $x \geq 1\ \ 2$

$$MCC_{AD,R} = MCC_{min} = \frac{np - u^2}{(1-x)xN^2} = \begin{matrix} p = (2x-1)N \\ n = 0 \\ u = n - xN \end{matrix} = \frac{x-1}{x}, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.24*

PRILOG 3.28 Izvod izraza za raspon vrijednosti parametra MCC za x $\leq 1\ \ 2$

$$\Delta(MCC_L) = MCC_{AA} - MCC_{AD,L} = 1 - \frac{x}{x-1} = \frac{1}{1-x}, \forall x \in [0, \frac{1}{2}]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.24*

PRILOG 3.29 Izvod izraza za raspon vrijednosti parametra MCC za $x \geq 1\ \ 2$

$$\Delta(MCC_R) = MCC_{AA} - MCC_{AD,L} = 1 - \frac{x-1}{x} = \frac{1}{x}, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.24*

Parametar *F1*

PRILOG 3.30 Pojednostavljenje parametra $F1$ za izmjenjive varijable ($o = u$)

$$F1 = \frac{2p}{2p + o + u} = \frac{2p}{2p + 2u} = \frac{p}{p + u}$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*, $F1$ –mjera točnosti predviđanja klase 1 kad se zanemari točnih predviđanja klase 0 (kojih je jako veliki broj).

PRILOG 3.31 Izvod izraza za maksimalnu vrijednost parametra $F1$

$$F1_{AA} = \frac{p}{p + u} = \frac{xN}{xN + 0} = 1, \forall x \in \langle 0,1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.31*

PRILOG 3.32 Izvod izraza za minimalnu vrijednost parametra $F1$ za $x \leq 1\ \ 2$

$$F1_{AD,L} = \frac{p}{p + u} = \frac{0}{xN + 0} = 0, \forall x \in \langle 0, \frac{1}{2}]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.31*

PRILOG 3.33 Izvod izraza za minimalnu vrijednost parametra $F1$ za $x\ \geq\ 1\ \ 2$

$$F1_{AD,R} = \frac{p}{p + u} = \frac{(2x - 1)N}{N(2x - 1) + (N - xN)} = \frac{2x - 1}{x}, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.31*

PRILOG 3.34 Izvod izraza za raspon vrijednosti parametra $F1$ za $x\ \geq\ 1\ \ 2$

$$\Delta(F1_R) = 1 - \frac{(2x - 1)}{x} = \frac{1 - x}{x}, \forall x \in [\frac{1}{2}, 1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.31*

Parametar κ

PRILOG 3.35 Izvod parametra κ za izmjenjive varijable ($o = u$)

$$\kappa = \frac{\frac{p+n}{N} - \frac{n+o \;\; n+u \;+\; p+o \;\; p+u}{N^2}}{1 - \frac{n+o \;\; n+u \;+\; p+o \;\; p+u}{N^2}}$$

$$\kappa = \; o = u \; = \frac{\frac{p+n}{N} - \frac{p+u\;^2 + \; u+n\;^2}{N^2}}{1 - \frac{p+u\;^2 + \; u+n\;^2}{N^2}}$$

$$\kappa = \frac{n^2 + 2nu - Nn + p^2 + 2pu - Np + 2u^2}{-N^2 + n^2 + 2nu + p^2 + 2pu + 2u^2} = \frac{np - u^2}{n+u \;\; p+u}$$

***** značenja kratica x, N, p, i n, objašnjena su ispod izvoda u *Prilogu 3.1*; $Q_{2,rnd}$- prosječna nasumična točnost; o – broj netočno predviđenih podataka klase 0, u – broj netočno predviđenih podataka klase 1; L je oznaka lijevog podintervala vrijednosti udjela klase 1 (x), tj. kada je $x \leq 1\;2$; R je oznaka desnog podintervala vrijednosti udjela klase 1 (x), tj. kada je $x \geq 1\;2$; κ - Cohenov kapa

PRILOG 3.36 Općenita supstitucijska tablica elemenata matrice pogrešaka za izvođenje minimalne (*AD* poredak) i maksimalne (*AA* poredak) karakteristične vrijednosti mjera kvalitete modela za parove općenitih ($u \neq o$) binarnih varijabli (klasifikacijskih varijabli s dvije klase)

	***AD* poredak**		***AA* poredak**	
	interval: $\boldsymbol{x + y \leq 1}$	interval: $\boldsymbol{x + y \geq 1}$	interval: $\boldsymbol{x \geq y}$	interval: $\boldsymbol{x \leq y}$
p	0	$((x+y)-1)N$	Ny	Nx
n	$N(1-(x+y))$	0	$N(1-x)$	$N(1-y)$
u	xN	$(1-y)N$	$N(x-y)$	0
o	yN	$(1-x)N$	0	$N(y-x)$

***** x – udio klase 1 u varijabli E; y – udio klase 1 u varijabli M; značenja ostalih kratica objašnjena su ispod izvoda u *Prilogu 3.6*

Ostali izvodi

PRILOG 3.37 Izvedeni izrazi za minimalne (*AD* poredak) i maksimalne (*AA* poredak) karakteristične vrijednosti mjera kvalitete modela za parove općenitih ($u \neq o$) binarnih varijabli (klasifikacijskih varijabli s dvije klase).

Parametar	***AD* poredak varijabli**		***AA* poredak varijabli**	
	interval: $\boldsymbol{x + y \leq 1}$	interval: $\boldsymbol{x + y \geq 1}$	interval: $\boldsymbol{x \geq y}$	interval: $\boldsymbol{x \leq y}$
Q_2	$1 - y - x$	$x + y - 1$	$y - x + 1$	$x - y + 1$
$Q_{2,rnd}$	$2xy - y - x + 1$	$2xy - y - x + 1$	$2xy - y - x + 1$	$2xy - y - x + 1$
ΔQ_2	$-2xy$	$-2(x-1)(y-1)$	$-2y(x-1)$	$-2x(y-1)$
MAE	$x + y$	$2 - y - x$	$x - y$	$y - x$
s	$\overline{(x+y)}$	$\overline{(2-y-x)}$	$\overline{(x-y)}$	$\overline{(y-x)}$
MCC	$-\overline{\frac{xy}{(1-x)(1-y)}}$	$\overline{\frac{(1-x)(1-y)}{xy}}$	$\overline{\frac{y(1-x)}{x(1-y)}}$	$\overline{\frac{x(1-y)}{y(1-x)}}$
κ	$-2\frac{xy}{x+y-2xy}$	$\frac{-2(x-1)(y-1)}{(x+y-2xy)}$	$\frac{-2y(x-1)}{x+y-2xy}$	$\frac{-2x(y-1)}{x+y-2xy}$

***** x – udio klase 1 u varijabli E; y – udio klase 1 u varijabli M; značenja ostalih kratica objašnjena su ispod izvoda u *Prilogu 3.6*

PRILOG 3.38 Izvod standardne devijacije σ pomoću x-a [83]

$$\sigma = \overline{\frac{1}{N-1} \sum_{i=1}^{N} (x - x^2)}$$

$$\{x = \mathrm{x}\}$$

$$\sigma = \overline{\frac{1}{N-1}\ xN\ 1-x\ ^2 + N\ 1-x\ \ 0-x\ ^2}\ , \forall x \in [0,1]$$

$$\sigma = \overline{xN\frac{(1-x)}{N-1}} \forall x \in [0,1]$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.35* , x – srednja vrijednost, x – u prvoj formuli bilo koji broj u varijabli, u drugoj formuli udio klase 1, σ – standardna devijacija

PRILOG 3.39 Izvod izraza za standardnu devijaciju eksperimentalne varijable σ_E pomoću udjela klase 1 (x) i elemenata matrice pogrešaka p, n, o i u

$$\sigma_E = \sqrt{xN\frac{(1-x)}{N-1}} \forall x \in [0,1]$$

$$\{x = \frac{p+u}{N}\}$$

$$\sigma_E = \sqrt{N\frac{\frac{p+u}{N}1-\frac{p+u}{N}}{N-1}}$$

$$\sigma_E = \sqrt{\frac{p^2}{N-N^2}+\frac{u^2}{N-N^2}+\frac{p}{N-1}+\frac{u}{N-1}+2p\frac{u}{N-N^2}}$$

$$\sigma_E = \sqrt{\frac{(p+u)(N-p-u)}{N(N-1)}} = \sqrt{\frac{(p+u)(n+o)}{N(N-1)}}$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6*, i *Prilogu 3.43*

PRILOG 3.40 Izvod izraza za standardnu devijaciju srednje vrijednosti eksperimentalne varijable SE pomoću udjela klase 1 (x) i elemenata matrice pogrešaka p, n, o i u

$$SE = \frac{\sigma}{\sqrt{N}}$$

$$SE = \frac{\sqrt{Nx\frac{(1-x)}{N-1}}}{\sqrt{N}} = \sqrt{x\frac{(1-x)}{(N-1)}}$$

$$SE = \sqrt{\frac{\frac{p+u}{N}(1+\frac{-p-u}{N})}{N-1}}$$

$$SE = \sqrt{\frac{(p+u)\frac{N-p-u}{N}}{N(N-1)}}$$

$$SE = \sqrt{\frac{(p+u)(N-p-u)}{N^2(N-1)}}$$

$$SE = \frac{1}{N}\sqrt{\frac{(p+u)(N-p-u)}{N-1}}$$

***** značenja svih kratica objašnjena su ispod izvoda u *Prilogu 3.6* i u *Prilogu 3.35*

Izvorni kodovi

PRILOG 3.41 Izvorni kodovi - Simulator

Simulator opisan u materijalima i metodama služi kako bi permutacijskim analizama bile provjerene formule izvedene u disertaciji za (1) maksimalnu, (2) minimalnu i (3) prosječnu nasumičnu vrijednost parametara kvalitete $Q_2, Q_{2,rnd}, \Delta Q_2, s, MAE, MCC, \kappa, F1$ (tj. mjere kvalitete) klasifikacijskog modela (u odnosu na eksperimentalnu vrijednost) prilagođene i primijenjene u analizi složenosti klasifikacijskih varijabli s dva stanja (s dvije vrijednosti, npr. samo 1 i 0). Pritom su: Q_2 *i* $Q_{2,rnd}$ parametri slaganja/točnosti klasifikacijskih varijabli E i M (ili dva klasifikacijska modela)

Simulator računa vrijednosti parametara kvalitete koji su inače definirani za procjenu točnosti modela $Q_2, Q_{2,rnd}, \Delta Q_2, s, MAE, MCC, \kappa, F1$ između stvarne varijable u njenom originalnom poretku (E, eksperimentalni/stvarni poredak) i modelne varijable (M, model) iste te varijable u nekom od presloženih/permutiranih poredaka.

Razlikujemo pritom tri slučaja:

(1) Kad su vrijednosti varijabli E i M identično poredane (uparene, tj. upareno sortirane) – pritom su vrijednosti parametara slaganja/korelacije ($Q_2, \Delta Q_2, MCC, \kappa, F1$) maksimalne a vrijedsnoti paramertare pogreške (S, MAE) minimalna (stoga što maksimalna korelacija odgovara minimalnoj pogrešci).

(2) Kad su vrijednosti varijabli E i M obrnuto poredane (obrnuto sortirane) - tj. kada su vrijednosti varijable E poredane silazno padajuće, a vrijednosti varijable M uzlazno rastuće, i

(3) Kad su vrijednosti varijable E i M u originalnom poretku varijabli i za svaki od poredaka računa.

Općenite formule koje se primjenjuju za procjenu kvalitete klasifikacijskih modela prilagođuju se za primjenu u permutacijskim analizama i analizama složenosti klasifikacijske varijable tako što se u njih uvede pojednostavljenje $u = o$,

Za svaki parametar nakon sortiranja računa se maksimalna i minimalna vrijednost što će kasnije služiti kao temelj analiza, i kao test za izvedene formule.

Za pokretanje simulacije u ovom projektu moguće je koristiti dvije skripte. Prva skripta je „simetrijskeStandalone.R", a druga „simetrijskeStandalonePar.R". Obje skripte izvršavaju istu funkciju, ali je razlika u načinu na koji se vrši paralelizacija. R skripta „simetrijskeStandalonePar.R" je paralelizirana verzija koda s automatiziranim odabirom omjera klasa u izmjenjivim varijablama. U slučaju da je potrebno paralelizirati verziju koja nije paralelna, to je moguće istovremenim višestrukim pokretanjem skripte „simetrijskeStandalonePar.R".

Prilikom opisa koda simulatora, bit će preskočene one linije koda koje nisu nužne za rad simulatora.

PRILOG 3.41 (a) Pokretanje aplikacije za simuliranje - simetrijskeStandalone.R

Ova skripta pokreće simulaciju procesa modeliranja i pri tome vraća rezultat u obliku CSV datoteke. Pokretanje „simetrijskeStandalone.R" skripte vrši se pomoću Rscript naredbe.

```
$ Rscript simetrijskeStandalone.R
Loading required package: data.table
Loading required package: ggplot2
Loading required package: labeling
Loading required package: FNN
Loading required package: Metrics
[1] "./program <EXP count> <MDL count> <1's in E or M> <length>"
[1] "e.g. expCnt=1000;  testCnt=100;  expOrTest1Cnt=50; len = 100"
```

Opis parametara aplikacije je slijedeći:

<EXP count> broj unikatnih vrijednosti eksperimentalne varijable

<MDL count> ukupan broj varijabli u modelu za svaku eksperimentalnu vrijednost varijable

<1's in E or M> broj klase 1 u eksperimentalnoj i modeliranoj varijabli

<length> veličina obiju varijabli (N)

Korišteno je $N = 100$ podataka za sve slučajeve pri čemu je broj eksperimentalnih varijabli 1000, $expCnt = 1000$, a za svaku od njih je generirano po 100 različitih modeliranih varijabli, $testCnt = 100$.

Primjer pokretanja simulatora pomoću koda u linux terminalu.

```
for x in `seq 1 99`; do
  Rscript simetrijskeStandalone.R  1000 100 $x 100 1> $x.1.txt 2>
$x.2.txt &
done
```

U navedenom primjeru izvršena je paralelizacija koda uzastopnim pokretanjem skripte „simetrijskeStandalone.R" pomoću Rscript R programa unutar „for petlje" u BaSh (Bourne Again Shell) jeziku [92]. Prvi parametar je ime skripte, a ostali parametri su parametri koji su opisani gornjem primjeru. Standardni izlaz aplikacije (stdout) preusmjeren je u datoteku $x.1.txt, dok je standardna pogreška aplikacije (stderr) preusmjeren u datoteku $x.2.txt pri čemu je „$x" promjenjiva vrijednost pa time nastaju datoteke naziva od „1.1.txt" do „99.1.txt" itd. Cijela linija je prebačena u pozadinu pomoću „&" znaka. Na taj način omogućen je istovremen rad 99 instanci programa.

Ova skripta radi sa R verzijom 3.6.X pa je zbog mogućih konflikata sa starijim verzijama R jezika postavljena staza za biblioteke u poseban folder (libPaths) [58].

```
      dir.create('~/R/x86_64-pc-linux-gnu-library/3.6', showWarnings = FALSE,
recursive = TRUE)
      .libPaths(new='~/R/x86_64-pc-linux-gnu-library/3.6')
      pkgTest <- function(x)
      {
        if (!require(x,character.only = TRUE))
        {
          install.packages(x,dep=TRUE)
          if(!require(x,character.only = TRUE)) stop("Package not found")
        }
      }
      pkgs = c("data.table","ggplot2","labeling","FNN","Metrics")
      for (pkg in pkgs){
        pkgTest(pkg)
      }
```

PkgTest [93] funkcija služi za provjeru je li paket (library) već instaliran, a ukoliko nije, instalira ga ili prekine s radom programa.

Nakon instaliranja potrebnih biblioteka, program učitava argumente unesene u komandnoj liniji.-Program se prekida ako nema dovoljno argumenata.

```
      args <- commandArgs(TRUE)
      if (length(args)<4){
        print("./program <EXP count> <MDL count> <1's in E or M> <length>")
        print("e.g. expCnt=1000;  testCnt=100;  expOrTest1Cnt=50; len = 100")
        stop()
      }
```

Nakon toga slijedi učitavanje svih datoteka programa koje sadrže izvorni kod i sve potrebne R biblioteke.

```
      source("./docLib.R")
      source("./docLib2.R")
      dl = DocLibClass$new()
```

```
library("data.table")
library("ggplot2")

source("./simulationsHskiki.R")
start.time <- Sys.time()
```

Datoteke docLib.R i docLib2.R sadrže izvorne kodove koje olakšavaju rad ove aplikacije. Biblioteka „data.table„ služi za učitavanje velikih tablica tamo gdje bi korištenje „data.frame“ bilo ograničavajuće rješenje. Biblioteka „ggplot2” služi za vizualizaciju podataka. Iako u ovom dijelu koda „ggplot2” nije neophodan ipak je ostavljen za potrebe debugiranja koda. Datoteka „simulationsHskiki.R” sadrži potrebnu logiku za pokretanje simulacija.

Za pokretanje simulacija potrebni su sljedeći parametri: broj jedinstvenih eksperimentalnih varijabli i broj testnih varijabli za svaku eksperimentalnu varijablu, a učitavaju se iz komandne linije koje skripta dobiva putem argumenata.

```
expCnt=as.integer(args[1])
testCnt=as.integer(args[2])
exp1Cnt=as.integer(args[3])
test1Cnt=exp1Cnt
len=as.integer(args[4])
```

Argumenti su objašnjeni u gornjem primjeru, a njihov redoslijed u kodu se podudara s-redoslijedom u komandnoj liniji. To su broj unikatnih eksperimentalnih varijabli, broj testnih/modeliranih varijabli za svaku eksperimentalnu varijablu, broj klase 1 u eksperimentu i ukupna duljina svake varijable bilo da je riječ o eksperimentalnoj ili modeliranoj varijabli.

Prilikom pokretanja programa, starta se timer na taj način da se trenutno vrijeme spremi u varijablu start.time. Nakon toga definira se naziv izlazne datoteke koristeći ulazne varijable.

```
start.time <- Sys.time()
outFile =
paste("out/simetric_expCnt",expCnt,"testCnt",testCnt,"exp1Cnt",exp1Cnt,"test1Cn
t",test1Cnt,"len",len,".csv", collapse = "_", sep = "")
if (!file.exists(outFile)){
  data=simulationsMultilen(expCnt,testCnt,exp1Cnt,test1Cnt,len)
  #data = dl$rotateMinMaxInTable(data)
  #data = dl$restoreMinMaxInTable(data)
  write.csv(data,file = outFile)
}
end.time <- Sys.time()
```

```
time.taken <- end.time - start.time
print(paste(exp1Cnt,time.taken, sep=" ", collapse = " "))
```

Ukoliko izlazna datoteka već ne postoji, pokreće se funkcija za obradu podataka „simulationsMultilen" s ulaznim parametrima, a njezini rezultati se spremaju u „data" varijablu, koja se kasnije spremi u izlaznu datoteku.

Po završetku spremi se trenutno vrijeme i usporedi s postojećim da se vidi duljina trajanja procesa. Poziv na funkciju „dl$rotateMinMaxInTable", koja se nalazi u „DocLib2" klasi, je zakomentiran, a on služi kako bi se krivi nazivi varijabli iskorigirali zbog sistematskog načina u imenovanju varijabli. Uzlazno-uzlazno poredane varijable označavaju se sufiksom Max, a uzlazno-silazno sortirane varijable (eksperimentalna i modelirana) označavaju sufiksom Min.

Funkcija „simulationsMultilen" nalazi se u zasebnoj datoteci „simulationsHskiki.R".

PRILOG 3.41 (c) Opis kod simulatora – jednostavni simulator - simulationsHskiki.R

U skripti „simulationsHskiki.R" nalazi se nekoliko vrsta simulacijskih funkcija od kojih će biti opisana samo jedna koja je ključna za rad simulatora, a riječ je o funkciji „simulationsMultilen".

```
simulationsMultilen=function(expCnt,testCnt,exp1Cnt,test1Cnt,len=100){
  tbl=data.table(
    exp=character()
    ,test=character()
    ,minCor=double()
    ,maxCor=double()
    ,natCor=double()
    , n=integer()
    , p=integer()
    , o=integer()
    , u = integer()
    , nMin=integer()
    , pMin=integer()
    , oMin=integer()
    , uMin = integer()
    , nMax=integer()
    , pMax=integer()
    , oMax=integer()
    , uMax = integer()
    , cntA = integer()
    , cntB = integer()
    , spearman = double()
```

```
    , spearmanMax = double()
    , spearmanMin = double()
    , kendall = double()
    , kendallMax = double()
    , kendallMin = double()
  )
```

Nakon definicije funkcije „simulationsMultilen" definira se varijabla tipa data.table (varijabla tbl) koji će držati sve podatke u memoriji da bi kasnije bili vraćeni return naredbom kao izlaz funkcije.

Nakon definicije tbl varijable slijedi stvaranje eksperimentalnih varijabli - onoliko koliko je definirano argumentom funkcije „expCnt".

```
  for (i in 1:expCnt){
    a = docLib.getRandomBinary(len,exp1Cnt)
    strA = paste(a,collapse="")
    sortedAa=docLib.sortBinary(a,TRUE)
    #sortedDa=docLib.sortBinary(a,TRUE)
```

U varijablu „a" sprema se niz slučajnih binarnih brojeva duljine definirane u „len" argumentu funkcije , nakon toga se taj niz pretvara u string i sprema u „strA" varijablu kako bi bio spremljen u datoteku kasnije. Generiranje slučajnih brojeva radi se funkcijom „docLib.getRandomBinary" iz „docLib.R" datoteke. Sortirana verzija originalnog niza „a" spremljena je u „sortedAa" varijablu po uzlaznom poretku. Funkcija „docLib.sortBinary" nalazi se u „docLib.R" datoteci i ona služi kako bi se brže sortiralo u odnosu na postojeće R implementacije sortiranja. Varijable „a", „strA", „sortedAa" odnose se na eksperimentalne varijable.

Isti se proces radi i za „b" ili „strB", „sortedDb", „sortedAb" varijable koje se odnose na modelirane (testne) varijable, a koje se ovdje generiraju za svaku eksperimentalnu varijablu i to onoliko različitih koliko je definirano u varijabli „testCnt".

```
    for (j in 1:testCnt){
      b = docLib.getRandomBinary(len,test1Cnt)
      strB = paste(b,collapse="")
      sortedAb=docLib.sortBinary(b,TRUE)
      sortedDb=docLib.sortBinary(b,FALSE)
      minCor=cor(sortedAa,sortedDb)
      maxCor=cor(sortedAa,sortedAb)
      natCor=cor(a,b)
      res = list(
        exp = strA
        , test = strB
        , minCor = minCor
        , maxCor = maxCor
```

```
            , natCor = natCor
            , n = docLib.getHitsNBin(a,b)
            , p = docLib.getHitsPBin(a,b)
            , o = docLib.getHitsOBin(a,b)
            , u = docLib.getHitsUBin(a,b)
            , nMin = docLib.getHitsNBin(sortedAa,sortedDb)
            , pMin = docLib.getHitsPBin(sortedAa,sortedDb)
            , oMin = docLib.getHitsOBin(sortedAa,sortedDb)
            , uMin = docLib.getHitsUBin(sortedAa,sortedDb)

            , nMax = docLib.getHitsNBin(sortedAa,sortedAb)
            , pMax = docLib.getHitsPBin(sortedAa,sortedAb)
            , oMax = docLib.getHitsOBin(sortedAa,sortedAb)
            , uMax = docLib.getHitsUBin(sortedAa,sortedAb)
            , cntA = exp1Cnt
            , cntB = test1Cnt
            , spearman = cor(a,b, method = "spearman")
            , spearmanMin = cor(sortedAa,sortedDb, method = "spearman")
            , spearmanMax = cor(sortedAa,sortedAb, method = "spearman")
            , kendall = cor(a,b, method = "kendall")
            , kendallMin = cor(sortedAa,sortedDb, method = "kendall")
            , kendallMax = cor(sortedAa,sortedAb, method = "kendall")

        )
        tbl=rbind(tbl, res)
      }
    }
```

Nakon stvaranja varijabala i njihovog sortiranja, računaju se korelacije i ostali parametri te spremaju u „res“ listu koja se kasnije nadodaje na kraj „tbl data.table“ varijable.

Metode „getHitsNBin”, „getHitsPBin”, „getHitsOBin”, „getHitsUBin” iz „docLib2.R“ datoteke (DocLibClass klasa) služe kako bi izračunale n, p , o i u parametre, bilo da je riječ o simulacijskim, minimalnim ili maksimalnim vrijednostima.

Izvan petlje su varijable koje se izračunavaju iz već izračunanih parametara.

```
    tbl$S=sqrt((tbl$u+tbl$o)/len)
    tbl$Q2=100*(tbl$p+tbl$n)/(tbl$p+tbl$n+tbl$o+tbl$u)
    tbl$Q2rnd = docLib.getQ2rnd(tbl$n, tbl$p, tbl$o, tbl$u)
    tbl$deltaQ = tbl$Q2-tbl$Q2rnd
```

```
        tbl$MCC = (tbl$n*tbl$p-tbl$u*tbl$o)/sqrt((tbl$p + tbl$o)*(tbl$p +
tbl$u)*(tbl$n+tbl$o)*(tbl$n+tbl$u));

        tbl$Mae = (tbl$o + tbl$u)/(tbl$p+tbl$n+tbl$o+tbl$u)

        tbl$cohenKappa=tbl$deltaQ/(1-tbl$Q2rnd)

        tbl$F1=(2*tbl$p)/(2*tbl$p+tbl$o+tbl$u)


        tbl$SMin=sqrt((tbl$uMin+tbl$oMin)/len)

        tbl$Q2Min=100*(tbl$pMin+tbl$nMin)/(tbl$pMin+tbl$nMin+tbl$oMin+tbl$uMin)

        tbl$Q2rndMin = docLib.getQ2rnd(tbl$nMin, tbl$pMin, tbl$oMin, tbl$uMin)

        tbl$deltaQMin = tbl$Q2Min-tbl$Q2rndMin

        tbl$MCCMin = (tbl$nMin*tbl$pMin-tbl$uMin*tbl$oMin)/sqrt((tbl$pMin +
tbl$oMin)*(tbl$pMin + tbl$uMin)*(tbl$nMin+tbl$oMin)*(tbl$nMin+tbl$uMin));

        tbl$MaeMin = (tbl$oMin +
tbl$uMin)/(tbl$pMin+tbl$nMin+tbl$oMin+tbl$uMin)

        tbl$cohenKappaMin=tbl$deltaQMin/(1-tbl$Q2rndMin)

        tbl$F1Min=(2*tbl$pMin)/(2*tbl$pMin+tbl$oMin+tbl$uMin)


        tbl$SMax=sqrt((tbl$uMax+tbl$oMax)/len)

        tbl$Q2Max=100*(tbl$pMax+tbl$nMax)/(tbl$pMax+tbl$nMax+tbl$oMax+tbl$uMax)

        tbl$Q2rndMax = docLib.getQ2rnd(tbl$nMax, tbl$pMax, tbl$oMax, tbl$uMax)

        tbl$deltaQMax = tbl$Q2Max-tbl$Q2rndMax

        tbl$MCCMax = (tbl$nMax*tbl$pMax-tbl$uMax*tbl$oMax)/sqrt((tbl$pMax +
tbl$oMax)*(tbl$pMax + tbl$uMax)*(tbl$nMax+tbl$oMax)*(tbl$nMax+tbl$uMax));

        tbl$MaeMax = (tbl$oMax +
tbl$uMax)/(tbl$pMax+tbl$nMax+tbl$oMax+tbl$uMax)

        tbl$cohenKappaMax=tbl$deltaQMax/(1-tbl$Q2rndMax)

        tbl$F1Max=(2*tbl$pMax)/(2*tbl$pMax+tbl$oMax+tbl$uMax)


        return(tbl)

      }
```

Rezultat funkcije se vraća naredbom return(tbl).

PRILOG 3.41 (d) Opis koda simulatora – datoteka s pomoćnim funkcijama - DocLib2.R

Ova datoteka sadrži klasu u kojoj je sadržana većina važnih pomoćnih metoda za obradu podataka u ovom radu. Pisana je u „R6" OOP-u i koristi „compiler" library za ubrzanje rada koda. U ovom dijelu bit će objašnjeno binarno sortiranje koristeći „sortBinary" metodu i generiranje binarnih brojeva koristeći „getRandomBinary" metodu.

```
library(R6)
library("compiler")


DocLibClass <- R6Class("DocLibClass", list())
DocLibClass$set("public", "sortBinary", cmpfun(function(arr,asc){
  if (is.list(arr)) arr=as.array(arr)
  sumArr=sum(arr)
  if (asc){
    arr2=c(rep(0,length(arr)-sumArr),rep(1,sumArr))
  } else {
    arr2=c(rep(1,sumArr),rep(0,length(arr)-sumArr))
  }
  return(arr2)
}))
```

Metoda „sortBinary" služi za sortiranje binarnih brojeva i radi na principu counting sort-a [94] Brojeve može sortirati uzlazno i silazno, ovisno o potrebi. Ta metoda se koristi u radu u svrhu ubrzanja sortiranja binarnih podataka u odnosu na postojeći algoritam implementiran u R jeziku za male nizove binarnih brojeva ($N < 10000$) gdje se pokazala brža od shell, quick i radix algoritama.

Sort radi na taj način da prebroji jedinice u nizu (varijabli) sumirajući ih i generira niz klase 1, one veličine koliko je te klase, a nakon toga generira broj 0, onoliko koliko je preostalo do duljine varijable. Spajajući ta dva niza dobiva se rezultat ovisno o odabranom redoslijedu.

Za testove i simulacije koristi se „getRandomBinary" metoda koja stvara niz napravljen od nula i jedinica.

```
DocLibClass$set("public", "getRandomBinary",
cmpfun(function(size,countOf1){
   p = countOf1/size
   x = rbinom(n=size, size=1, prob=p)
   if (sum(x) != countOf1){
     return(self$getRandomBinary(size,countOf1))
   }
   return(x)
}))
```

Funkcija u svojem radu generira niz pomoću binomne razdiobe, provjerava ima li dovoljno jedinica u njoj. Ukoliko nema dovoljno jedinica, poziva samu sebe sve dok se uvjet ne zadovolji.

Test brzine rađen je na koristeći microbenchmark [95] R biblioteku:

```
      sortBenchmark=function(arr){
        res=list()
        res["quick"]=mean(microbenchmark(  sort.int(arr,
method="quick")  )$time/1000)
        res["shell"]=mean(microbenchmark(  sort.int(arr,
method="shell")   )$time/1000)
        res["sort"]=mean(microbenchmark(  sort(arr)  )$time/1000)
        res["radix"]=mean(microbenchmark(  sort.int(arr,
method="radix")  )$time/1000)

res["sortBinary"]=mean(microbenchmark(  dl$sortBinary(arr,TRUE)  )$time/1000)
        return (res)
      }


      sortBenchmark(arr = dl$getRandomBinary(100,10))
      sortBenchmark(arr = dl$getRandomBinary(100,90))
      sortBenchmark(arr = dl$getRandomBinary(1000,10))
      sortBenchmark(arr = dl$getRandomBinary(1000,15))
      sortBenchmark(arr = dl$getRandomBinary(1000,100))
```

Rezultati testa brzine koda su slijedeći:

```
      > sortBenchmark(arr = dl$getRandomBinary(100,10))
      $quick
      [1] 22.50131


      $shell
      [1] 21.08848


      $sort
      [1] 53.82768


      $radix
      [1] 41.79104


      $sortBinary
```

```
[1] 6.27771

> sortBenchmark(arr = dl$getRandomBinary(1000,10))
$quick
[1] 33.90265

$shell
[1] 29.38238

$sort
[1] 67.88019

$radix
[1] 46.78309

$sortBinary
[1] 12.89776

> sortBenchmark(arr = dl$getRandomBinary(1000,15))
$quick
[1] 38.40981

$shell
[1] 29.74066

$sort
[1] 53.70515

$radix
[1] 45.74966

$sortBinary
[1] 12.89127

> sortBenchmark(arr = dl$getRandomBinary(1000,100))
$quick
[1] 53.61087

$shell
```

```
      [1] 31.50064

      $sort
      [1] 73.18506

      $radix
      [1] 51.59688

      $sortBinary
      [1] 12.67778
```

Vrijeme izvršavanja se pokazalo za „sortBinary" manje u svim testnim slučajevima malenih nizova pa je poziv na metodu „sortBinary" integriran u simulator.

Slijedeće metode služe za izračunavanja parametara koji sadrže informaciju o broju i vrsti pogodaka/pogrešaka, bilo da se radi o parametrima n, p, o ili u.

```
      DocLibClass$set("public", "getHitsPBin",
cmpfun(function(experimentalArr,modelArr){
        N=length(experimentalArr)
        if (N != length(modelArr)) return(NA)
        x=0
        for (i in 1:N){
          if (experimentalArr[i]*modelArr[i]==1){
            x=x+1
          }
        }
        return(x)
      }))


      DocLibClass$set("public", "getHitsNBin",
cmpfun(function(experimentalArr,modelArr){
        N=length(experimentalArr)
        if (N != length(modelArr)) return(NA)
        x=0
        for (i in 1:N){
          if (experimentalArr[i]+modelArr[i]==0){
            x=x+1
          }
        }
        return(x)
```

```
      }))
      DocLibClass$set("public", "getHitsUBin",
cmpfun(function(experimentalArr,modelArr){
        N=length(experimentalArr)
        if (N != length(modelArr)) return(NA)
        x=0
        for (i in 1:N){
          if (experimentalArr[i]>modelArr[i]){
            x=x+1
          }
        }
        return(x)
      }))
      DocLibClass$set("public", "getHitsOBin",
cmpfun(function(experimentalArr,modelArr){
        N=length(experimentalArr)
        if (N != length(modelArr)) return(NA)
        x=0
        for (i in 1:N){
          if (experimentalArr[i]<modelArr[i]){
            x=x+1
          }
        }
        return(x)
      }))
```

PRILOG 3.41 (e) Analize rezultata simulatora

Analize rezultata se pokreću pomoću dviju skripti „analizeSvihPodataka.R" i „analizeRaspona.R". Samo datoteke koje su nužne za dobivanje rezultata bit će opisane. Sve ostale su nebitne i služile u samo u nekom trenutku prilikom razvoja softvera i raznih analiza.

Datoteke iz ovog poglavlja moguće je pokrenuti direktno iz Rstudio alata [64]. Izvršavanje se vrši po potrebi liniju po liniju, jer parametarski odabir nije napravljen u programu tako da su linije, koje se u nekom trenutku ne koriste, zakomentirane znakom „#".

Analize rezultata simulatora - Analize svih podataka (analizeSvihPodataka.R)

U datoteci „analizeSvihPodataka.R" radi se analiza svih podataka koji su dobiveni simulacijama. Simulacije su rađene u postocima od 1 do 99 % klase 1 u vektoru. Svaka datoteka sadrži 1000 slučajnih eksperimentalnih varijabli od kojih svaka ima po 100 slučajnih modela.

U prvom dijelu datoteke „analizeSvihPodataka.R” pozivaju se potrebne biblioteke i izvorni kodovi koji su potrebni za nesmetan rad aplikacije.

```
source("./docLib2.R")

#install.packages("data.table")
#install.packages("foreach")
#install.packages("doParallel")

library(foreach)
library(doParallel)
library(sqldf)

library("data.table")
library("ggplot2")
folder = "out/"
files = list.files(folder,  pattern="simetric.*.csv")
```

Izlazna mapa se postavlja u folder varijabli i na toj lokaciji će se stvarati svi izlazni tablični ili tekstualni rezultati.

Nakon istanciranja DocLibClass klase (objekt dl, klasa učitana iz DocLib2.R datoteke ranije spomenute u dijelu gdje je objašnjen simulator) slijedi učitavanje postojećih podataka. U direktoriju „analyzeAll” nalaze se datoteke koje se odnose na analizu negrupiranih podataka. Taj direktorij sadrži R izvorne kodove koji se po potrebi pozivaju „source“ naredbom.

```
 dl = DocLibClass$new()
res= list()

#create big data
#big data loader
source("inc/analyzeAll/bigDataLoader.R")
source("inc/analyzeAll/addExtraVars.R")

source("inc/analyzeAll/entropyCalculator.R")
#write.csv(bigData,"out/bigData.csv")
```

U danom gornjem primjeru podaci su učitani pomoću „bigDataLoader.R” datoteke (potprograma). Ta datoteka provjerava postoji li velika datoteka „out/bigData.csv” ili njena komprimirana verzija „out/bigData.csv.gz” i ukoliko postoji, učita je u data.table instancu objekta („bigData” varijabla).

Ukoliko ne postoji, uzima rezultate simulacije iz „out/" direktorija i povezuje ih sve u jednu datoteku „out/bigData.csv".

Potprogram „addExtraVars.R" postoje li sve varijable i dodaje varijable koje nisu automatski generirane prilikom simuliranja.

Nakon što su sve varijable dodane i podaci učitani, slijedi daljnja obrada podataka. Sve obrade u datoteci „analizeSvihPodataka.R" rade s varijablom „bigData".

```
source("inc/analyzeAll/entropyCorelator.R")
source("inc/analyzeAll/entropyCorelatorByXRange.R")
source("inc/analyzeAll/entropyVisualizer.R")
```

„entropyCorelator.R" iz bigData objekta korelira sve varijable povezane s entropijom i korelira ih s ostalim varijablama u svrhu detekcije koja varijabla je najsličnija entropiji. Rezultat analize sprema se u "out/correlationsWithEntropyALL.csv" datoteku

„entropyCorelatorByXRange.R" korelira polovice segmenta omjera klase 1 i ukupne duljine varijable (x) raznih parametara s entropijom. Rezultat analize sprema se u "out/entropyCorreltorByXRange.csv" datoteku

„entropyVisualizer.R" vizualizira odnos entropije i ostalih varijabli (scatter plot) i sprema slike u direktorij „plot/entropyVisualizer/".

```
source("inc/analyzeAll/rmseParamsVsEnt.R")


source("inc/analyzeAll/makeAggregations.R")
source("inc/analyzeAll/makeAggregationsMinMaxAvg.R")
source("inc/analyzeAll/makeAggregationsAvgStdev.R")
source("inc/analyzeAll/checkMinMaxFormulae.R")
```

„rmseParamsVsEnt.R" normalizira parametre te uspoređuje njihovu sličnost entropiji. Rezultat sprema u „out/rmseParamsVsEnt.csv". Analiza se vrši na vrijednostima simulacijskim, apsolutnim i rasponima

„makeAggregations.R" , „makeAggregationsMinMaxAvg.R" i „makeAggregationsAvgStdev.R" agregiraju podatke kako bi iz simulacijskih i apsolutnih vrijednosti računali agregacije, bilo da je riječ o minimalnim, maksimalnim ili vrijednostima raspona. Rezultati se spremaju u datoteke: „out/simetrijske_statFajlovaRASPON.csv", „out/simetrijske_statFajlovaRASPON_full.csv", „out/makeAggregationsAvgStdev.csv", „out/makeAggregationsMinMaxAvg.csv"

„checkMinMaxFormulae.R" uspoređuje stvarne podatke (apsolutne) za $Q_2, Q_{2,rnd}, \Delta Q_2, MAE, F1, s$ s onima dobivenim pomoću formula koristeći rmse funkciju.

Analize rezultata simulatora - Analize raspona podataka

Sve analize grupiranih podataka rade se kroz glavnu skriptu „analizeRaspona.R". Ta skripta poziva druge skripte koje se izvršavaju po potrebi na taj način da korisnik sam komentira one linije koje ne želi koristiti pomoću RStudio.

```
      source("inc/env.R")
        library(scales)


        library("Hmisc")
        source("inc/analyzeRange/dataLoader.R")


        cols=dl$getVarBases(colnames(data))
        dataVertical = dl$pivotTableToVertically(data,"percentage")
        dataVerticalMinMax =
dl$pivotTableToVertically(dataMinMax,"percentage")
```

U prvih nekoliko linija učitavaju se sve potrebne biblioteke i podaci iz CSV datoteka koje su dobivene obradom svih podataka. Datoteke s izvornim kodom se učitavaju iz direktorija „inc/analyzeRange".

Kad se podaci učitaju, rade se manipulacije s podacima zbog lakšeg grupiranja tako da se tablica s velikim brojem kolona pretvori u tablicu s tri kolone koje sadrže naziv kolone, vrijednost kolone i vrijednost one kolone u odnosu na koju se podaci organiziraju (npr. „x" ili „$percentage$").

```
        skipFields =
c("fileName","percentage","cntA","cntB","p","o","n","u")


        for (col in cols){
          if ( col %in% skipFields )  next()
          colX = paste("abs_delta_",col,sep = "", collapse = "")
          colY = paste("sim_delta_",col,sep = "", collapse = "")


          source("inc/analyzeRange/visualize_scatter_ABS_SIM.R")
          source("inc/analyzeRange/visualize_scatter_perc_abs_sim.R")
          source("inc/analyzeRange/visualize_scatter_var1var2.R")
          source("inc/analyzeRange/visualize scatter perc MinMaxAbsSim.R")
        }
```

U dijelu koda iznad, postavljena su polja za koja se obrade neće vršiti. Nakon toga slijedi petlja, koja iterira kroz sve kolone i za svaku kolonu koja nije u listi kolona za preskakanje, vrše se obrade. Nazivi kolona abs_delta i sim_delta odnose se na prefikse kolona simulacijske i apsolutne vrijednosti raspona varijabli.

visualize_scatter_ABS_SIM.R je potprogram koji generira slike u direktorij „plot/scatter_ABS_SIM", a te slike se odnose na simulacijske i apsolutne raspone

visualize_scatter_perc_abs_sim.R, stvara slike u direktoriju „plot/scatter_perc_abs_sim" koje predstavljaju odnos između parametara i x-varijable (udio klase 1 u ukupnoj duljini varijable)

visualize_scatter_var1var2.R je potprogram za vizualizaciju odnosa varijabli (scatter plot) i sprema ih u direktorij „plot/scatter_var1var2". Radi tako da za dvije apsolutne ili dvije simulacijske varijable napravi scatter plot

visualize_scatter_perc_MinMaxAbsSim.R vizualizira apsolutne i simulacijske vrijednosti minimuma i maksimuma na jednoj slici za svaku varijablu. Na y osi je odabrana varijabla, dok je na x osi „x" varijabla. Na grafovima je prikazana svaka druga točka zbog lakšeg prikaza. Slike su spremljene u direktorij „plot/scatter_perc_MinMaxAbsSim".

Za ispitivanje rubnih uvjeta izvedenih formula (ili regresijom dobivenih formula) koristi se varTester.R.

```
source("inc/analyzeRange/varTester.R")
```

„varTester.R" svoj izlaz sprema u „out/ranges/varTester.csv" i služi~~o~~ ~~je~~ samo prilikom provjere formula.

Slijedeći potprogrami provjeravaju korelacije među različitim varijablama:

```
source("inc/analyzeRange/korelacije_kolona_s_kombinacijama.R")
source("inc/analyzeRange/correlationWithEntBySegment.R")
source("inc/analyzeRange/korelacije_kolona_s_entropijom.R")
source("inc/analyzeRange/korelacije_kolona_s_entropijom_minMax.R")
source("inc/analyzeRange/minMaxCorrelatorByX.R")
```

Svaka od obrada izlazne rezultate sprema u CSV format datoteke u „out" direktoriju.

- korelacije_kolona_s_kombinacijama.R korelira sve datoteke s brojem kombinacija (comb, logComb kolone u data varijabli) i sprema u datoteku „korelacije_kolona_s_kombinacijama.csv"
- correlationWithEntBySegment.R - korelira varijable entropije sa segmentima x-a ($x < 0.5$ i $x \geq 0.5$)
- korelacije_kolona_s_entropijom.R - korelira entropije s ostalim kolonama raspona
- korelacije_kolona_s_entropijom_minMax.R - korelira entropije s ostalim kolonama minimuma i maksimuma
- minMaxCorrelatorByX.R - korelira minimalne i maksimalne vrijednosti s udjelom klase 1 podijeljene na dva segmenta ($x < 0.5$ i $x \geq 0.5$) i sprema u datoteku „out/minMaxCorreltorByX.csv"

Za provjeru izvedenih formula raspona koristi se slijedeći potprogram:

```
#source("inc/analyzeRange/provjeraVarPoFormulama.R")
```

S obzirom da su sve formule definirane kao funkcije x-a, posebno se provjerava lijeva strana ($x \leq 0.5$), a posebno desna ($x \geq 0.5$). Provjera se vrši tako da se sumiraju sve razlike u odnosu na stvarnu vrijednost raspona dobivenoga pomoću apsolutnih vrijednosti.

Kod vizualizacije scatter plotova važne su dva potprograma koji povezuju binarnu entropiju ($entX$) s ostalim varijablama.

```
source("inc/analyzeRange/visualize_scatter_entropy.R")
source("inc/analyzeRange/visualize_scatter_entropyAll.R")
```

Razlika između „visualize_scatter_entropy.R" i „visualize_scatter_entropyAll.R" je u tome što „All" verzija vizualizira odnose varijabli kombinacija (logaritmirane i nelogaritmirane) , p2, p2Log i entX sa svakom od ostalih kolona pojedinačno (i minimumi i maksimumi i rasponi). Izlaz se sprema u direktorij „plot/scatter_entropy/all".

„visualize_scatter_entropy.R" vizualizira odnos $entX$ (aproksimacije entropije pomoću stirlingove formule) s apsolutnim i simulacijskim vrijednostima raspona za svaku varijablu pojedinačno i prikazuje ih na istom grafu. Rezultat se nalazi u „plot/scatter_entropy" direktoriju:

```
if (!dir.exists("out/ranges")) mkdirs("out/ranges")
# source("inc/analyzeRange/dataGeneratorWithFormula.R")
# source("inc/analyzeRange/dataGeneratorWithFormulaMinMax.R")
```

Data generatori u ovom dijelu služili su samo za generiranje podataka koristeći formule za raspone.

PRILOG 3.42

U ovom dijelu će biti prikazani dijelovi koda koji su korišteni u simulacijskim analizama radi dobivanja karakterističnih (minimalnih, maksimalnih i prosječnih) vrijednosti parametara kvalitete klasifikacijskih modela prilagođeni u disertaciji analizi složenosti varijabli. Svi izvorni kodovi nalaze se u direktoriju „hskiki_ponovno".(*Prilog E_35*) docLib.R - datoteka s pomoćnim funkcijama za izračunavanje parametara

simulationsHskiki.R – logika za izračunavanje parametara

test11.R – program koji služi izvođenju simulacije u kojoj i eksperimentalna i modelirana (testna) varijabla imaju 50 % klase 1. Rezultati se spremaju u „out/plot/", a datoteke imaju prefiks test11

test12.R – program koji služi izvođenju simulacije u kojoj i eksperimentalna i modelirana (testna) varijabla imaju 80 % klase 1. Rezultati se spremaju u „out/plot/", a datoteke imaju prefiks test12

Primjer koda za test12 koji ima 80 % klase 1 u obje varijable počinje s učitavanjem docLib.R datoteke s izvornim kodom, poziva biblioteke data.table za velike tablice, te ggplot2 za vizualizaciju grafova (scatter plot i histogrami).

```
source("docLib.R")
library("data.table")
library("ggplot2")
source("./simulationsHskiki.R")
```

U slijedećim linijama definiraju se varijable, te (ukoliko već postoje) učitavaju se rezultati obrade. Ukoliko ne postoje, tada se mjeri vrijeme i na temelju varijabli (ulaznih parametara) vrši se obrada i sprema u data12 varijablu.

```
         expCnt=1000
        testCnt=100
        exp1Cnt=80
        test1Cnt=80
        if (file.exists("out/test12.csv")){
          data12=read.csv("out/test12.csv")
        }else{
          start.time <- Sys.time()
            data12=simulationsHskiki(expCnt,testCnt,exp1Cnt,test1Cnt)
          end.time <- Sys.time()
          time.taken <- end.time - start.time
          time.taken
          write.csv(data12,file = "out/test12.csv")
          if (!file.exists("out/test12_summary.csv"))
            write.csv(summary(data12),file = "out/test12_summary.csv")
        }
```

Obrada se vrši pomoću „simulationsHskiki" funkcije tako da joj se proslijede gore ranije definirane varijable (expCnt, testCnt, exp1Cnt, test1Cnt) kao argumenti. Po završetku obrade rezultat se sprema u datoteke „out/test12.csv" i „out/test12_summary.csv".

U linijama ispod obrade podataka, slijedi vizualizacija histograma. Definira se varijabla „binShift" koja osigurava da pomak grafa ostane prikazan u okvirima između linija minimuma i maksimuma kako apsolutnog (zelene linije), tako i simulacijskog (crvene linije).

```
      binshift =2
      plotQ2=ggplot(data12) +
        geom_histogram(aes(data12$Q2),
                       binwidth = 2, fill = "blue", color = "black")+
xlim(0,100) +
        ggtitle("Distribution of Q2 model quality parameter")+
        labs(y="Broj slučajeva/simulacija", x = "Q2")  +
        geom_vline( aes(xintercept=min(data12$Q2)-binshift),color='red')  +
        geom_vline( aes(xintercept=max(data12$Q2)),color='red') +
        geom_vline( aes(xintercept=min(data12$Q2Min)-
binshift),color='green' ,linetype="dashed")  +
        geom_vline( aes(xintercept=max(data12$Q2Min)-
binshift),color='green' ,linetype="dashed")  +

geom_vline( aes(xintercept=min(data12$Q2Max)),color='green' ,linetype="dashed")
+

geom_vline( aes(xintercept=max(data12$Q2Max)),color='green' ,linetype="dashed")
+
```

```
        geom_vline( aes(xintercept=mean(data12$Q2)),color='yellow')
      #theme(plot.title = element_text(size=9))


      ggsave("out/plot/test12-Q2.png",plotQ2,width = 12, height = 8, units =
"cm")
```

U gornjem primjeru je prikazan primjer koda vizualizacije histograma koristeći „ggplot2“ biblioteku. Vizualizacija je spremljena u „out/plot” direktorij.

Za prikaz scatter plota prikazan je primjer vizualizacije odnosa korelacije i ΔQ_2.

```
      plotPearsonDeltaQ=ggplot(data=data12, aes(x=deltaQ, y=natCor)) +
        geom_point(size=2, shape=21)+
        labs(y="Pearsonov koeficijent korelacije (R)" )+
        xlab(     expression("Doprinos modela "*Delta*Q[2])   )+
        geom_smooth(method=lm, se=FALSE)
      ggsave("out/plot/test12-PearsonDeltaQ.png",plotPearsonDeltaQ,width = 12,
height = 8, units = "cm")
```

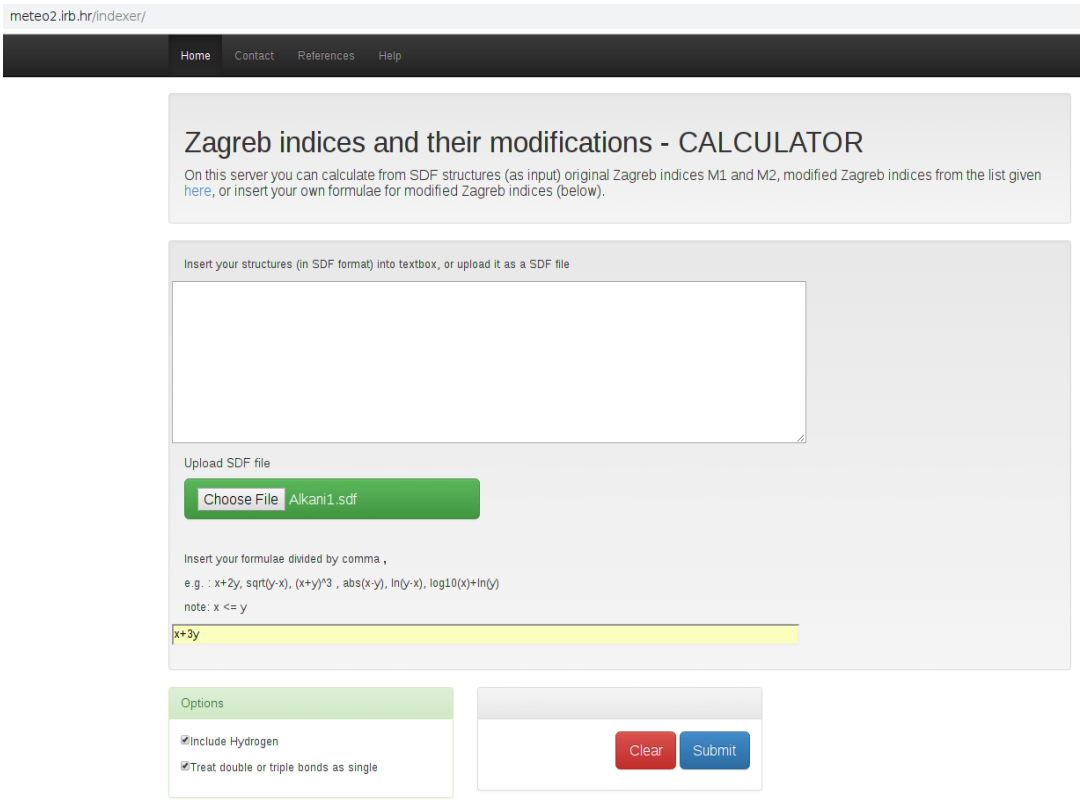
U primjeru se vizualizacija scatter grafa sprema u „out/plot” direktorij u „test12-PearsonDeltaQ.png” datoteku.

Na istom principu funkcionira i test11, pa će njegovo objašnjenje koda biti preskočeno.

PRILOG 3.43 Mrežni poslužitelj za izračun Zagrebačkih indeksa i njihovih modifikacija („*Zagreb indices and their modifications – CALCULATOR*“)

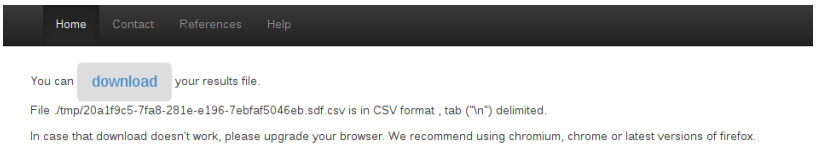
Ova aplikacija služi za izračunavanje različitih deskriptora iz kemijskih struktura dobivenih iz MOL/SDF datoteka. Njezin izvorni kod moguće je naći na slijedećoj adresi: https://github.com/vbojovic1980/sci-rep , a dostupna je javno na adresi http://meteo2.irb.hr/indexer/ odakle ju je moguće pokretati ili skinuti za rad na računalima.

Deskriptori se računaju iz kemijske strukture prebrojavajući valencije duljine 1 i 2, uz mogućnost filtriranja vodika, te tretiranja višestrukih veza kao jednostruke.



Koristeći oznake x i y (početna i krajnja vrijednost valencije veza kod veza dvaju atoma) moguće je unijeti i dodatne formule koje definiraju sami korisnici.

Nakon što su podaci poslani na server naredbom submit, pojavljuje se poruka slična slici ispod, s mogućnošću downloada rezultata.



Obrađeni podaci spremljeni su u datoteci u CSV formatu.

Prvih nekoliko linija odnosi se na parametre koji su korišteni da bi se ova analiza dobila. Nakon toga slijede linije s kolonama rezultata.

stupac 1 = redni broj molekule u MOL/SDF datoteci

stupac SMILES– kemijska formula molekule u SMILES formatu

[1,1]..[7,7] valencije veza i njihov ukupan broj u molekuli

F1..F17 izvorni i modificirani Zagrebački indeksi (deskriptori) izračunani za svaku molekule

Način na koji se računaju ti deskriptori te frekvencije veza različite vrste opisan je u literaturi – u objavljenom radu [53].

PRILOG 3.44 Mrežni poslužitelj za izračun složenosti varijabli

Classification variable complexity parameter estimator je web aplikacija napravljena, u sklopu ovog rada, u svrhu određivanja kompleksnosti podataka. Pomoću web aplikacije za ulaznu datoteku određuje se kompleksnost i ostale parametre za svaki stupac, pod uvjetom da su podaci u datoteci sastavljeni samo od podataka klase 0 i 1.

Server aplikacije je postavljen na adresu [76].

Projekt se sastoji od slijedećih datoteka:

DCAClass.R – klasa s logikom programa

docLib2.R – pomoćna klasa

env.R - potprogram za učitavanje biblioteka i njihovu automatsku instalaciju

DCAConsole.R – konzolna verzija ove aplikacije za slučajeve kada je prevelika količina podataka za obradu

DCAWeb.R – sučelje i logika web aplikacije

server.r – datoteka za učitavanje servera

ui.r - datoteka za učitavanje sučelja

(a) Logika web aplikacije - DCAClass.R

„DCAClass.R" datoteka sadrži „DCAClass" klasu koja je namijenjena izračunu svih potrebnih parametara za web i konzolnu aplikaciju. Klasa je napisana koristeći R6 OOP (biblioteka za korištenje objektno orijentiranog programiranja u R-u).

U definiciji klase „DCAClass" instancira se „DocLibClass" klasa u varijablu dl, te se postavlja fileName varijabla koja sadrži naziv datoteke koja se preko web aplikacije predaje serveru.

```
DCAClass <- R6Class("DCAClass", list(
  dl = DocLibClass$new()
  , res = list()
  , fileName = character()
))
```

Run metoda pokreće program.

```
DCAClass$set("public", "run",function(fileName,df = NULL){
```

Za pokretanje programa potrebna je CSV datoteka ili data frame / data table objekt iz kojeg će se preuzeti binarni brojevi.

Nakon pokretanja aplikacije run metodom, podatke je moguće eksportirati funkcijom „exportXLSX".

```
DCAClass$set("public", "exportXLSX", function(){
```

Funkcija „exportXLSX" sprema rezultate u XLSX format, na način da stvara novi list za svaku od tablica (simulacije i formule za retke i stupce). Ime izlazne xlsx datoteke ovisi o imenu datoteke koja je dana aplikaciji u run metodi.

Prilikom rada aplikacije za eksport podataka koristi se „getFileName" metoda.

```
DCAClass$set("public", "getFileName", function(isRow ,
isSimulation ,EXT="csv", fileName=NULL){
```

Metoda kao argumente uzima informaciju o tome hoće li biti simulacijom (isSimulation) ili formulom izračunati podaci u tablicama. U skladu s time podaci se eksportiraju u različite datoteke.

Primjeri izlaznih datoteka su slijedeći: „out.csv.col_simulation.csv", „out.csv.col_formula.csv". Sve te datoteke nastale su obradom datoteke „out.csv".

Nakon obrade svih podataka, a za potrebe web aplikacije, korisnik može dobiti rezultate u obliku zip datoteke. Naziv te datoteke dobiva se metodom „getZipFileName":

```
DCAClass$set("public", "getZipFileName", function( ){
```

U svrhu komprimiranja datoteke, potrebno je pokrenuti „zipIt" metodu.

```
DCAClass$set("public", "zipIt", function( ){
```

Metoda „zipIt" sprema sve CSV datoteke koje su rezultat obrade u jednu ZIP datoteku.

Metoda za eksport rezultata u CSV format se zove „export" metoda.

```
DCAClass$set("public", "export", function(){
```

U „export" metodi rezultati obrade iz „self$res" svojstva klase spremaju se u zasebne CSV datoteke kojima je prefiks (naziv ulazne datoteke) skupa sa sufiksom određen „getFileName" metodom (ranije objašnjenom).

Za potrebe eksporta u JSON format, a u svrhu čitanja iz JavaScript jezika, napisana je „getJson" metoda.

```
DCAClass$set("public", "getJson", function(){
```

Pomoćna metoda „factorsToStrings" koristi se za potrebe ispravljanja pogrešno postavljenih tipova koloni.

```
DCAClass$set("public", "factorsToStrings", function(df){
```

Unutar „run" metode poziva se metoda „doCalculations" koja obrađuje podatke data.frame ili data.table tablice. Ova metoda poziva obradu koja koristi izvedene formule, ali i vrši simulacije koristeći ranije formule koje koriste p, n, o, u parametre i sortiranja.

```
DCAClass$set("public", "doCalculations", function(data){
```

Za računanje minimalnih, maksimalnih i vrijednosti raspona, koristi se „calculateRowRanges" metoda. Pri tom postupku koriste se formule koje sadrže udio klase 1 u varijabli (x) kao argument jednadžbi .

```
DCAClass$set("public", "calculateRowRanges",function(a,rowNum,colName,
roundDigits=NULL){
```

Metoda „calculateRowRanges" računa raspone za nizove dane „a" argumentom. Osim tog argumenta, metodi je dan redni broj kolone, naziv kolone i broj decimalnih mjesta za zaokruživanje.

Za određivanje vrijednosti simulacijama, koristi se metoda „calculateRowSimulation" koja uzima iste argumente kao i „calculateRowRanges" s razlikom u načinu rada.

```
DCAClass$set("public",
"calculateRowSimulation",function(a,rowNum,colName, roundDigits=NULL){
```

Prilikom simulacija koristi se ubrzani algoritam za sortiranje binarnih brojeva i radi se na sličan način kako su rađene simulacije u pod-poglavlju o simulacijama, s tom razlikom da se ovdje simulacija odnosi na svaki stupac korisničkih podataka, a ne izmišljenih slučajnih podataka.

(b) Datoteka s pomoćnim funkcijama - docLib2.R

Klasa docLib2.R potrebna je za normalan rad aplikacije. Riječ je o istoj datoteci koja se spominje kroz ostale dijelove priloga ovog rada.

(c) Datoteka za učitavanje pomoćnih varijabli - env.R

Potprogram env.R služi za potrebe učitavanja biblioteka potrebnih za rad ovog programa. Ranije je već objašnjen kod analiza rezultata simulatora i radi na istom principu.

(d) Skripta za rad iz konzolne linije - DCAConsole.R

Verzija aplikacije DCAConsole.R može se pokrenuti lokalno na PC-u ili nekom drugom računalu koji ima instaliran R. Verzija R-a mora biti veća ili jednaka 3.6.0. Aplikaciju se pokreće Rscript naredbom na način prikazan na primjeru ispod (s ili bez time naredbe).

Argument koji aplikacija uzima je samo naziv datoteke i bez toga ne može raditi.

```
ezop@kanta:~/radni/irb/projekti/CA$ Rscript DCAConsole.R
[1] "Rscript DCAConsole.R <file.csv>"
```

Ukoliko se aplikaciji dade ispravna datoteka, tada će po završetku pokretanja ispisati nazive svih izlaznih datoteka, koje sadrže rezultate simulacija i izračuna pomoću formula za svaki od parametara.

```
ezop@kanta:~/radni/irb/istrazivanja/CA_cca$ time Rscript DCAConsole.R
test/test1000x100.csv
── Attaching packages ─────────────────────────────────── tidyverse 1.3.0 ──
✔ ggplot2 3.3.0     ✔ purrr   0.3.3
✔ tibble  2.1.3     ✔ dplyr   0.8.5
✔ tidyr   1.0.2     ✔ stringr 1.4.0
✔ readr   1.3.1     ✔ forcats 0.5.0
── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
[1] "test/test1000x100.csv"
[1] "test/test1000x100.csv.col_formula.csv"
[1] "test/test1000x100.csv.col_simulation.csv"


real    0m3.929s
user    0m3.788s
sys     0m0.112s
```

(e) Skripta za učitavanje web aplikacije - DCAWeb.R

U „DCAWeb.R" datoteci sadržana je web aplikacija pisana u R jeziku korištenjem Shiny web application framework for R [62].

U početnom dijelu programa, web aplikacija učitava izvorne kodove potrebne za svoj rad.

```
source("./DCAServer.R")
source("DCAUI.R")
```

„DCAServer.R" datoteka sadrži izvorni kod za definiciju servera

„DCAUI.R" datoteka sadrži izvorni kod za definiciju komponenti korisničkog sučelja.

Pokretanje poslužitelja vrši se u slijedećim linijama.

```
interactive()
setOption("shiny.port",2020)
shinyApp(ui = ui, server = server , options =getOption("shiny.port"))
```

Postavlja se interaktivni način rada, port se postavlja na 2020 i pokreće aplikacija na tom portu s postavljenim korisničkim sučeljem „ui" i poslužiteljem „server".

(f) Definicija korisničkog sučelja - DCAUI.R

Biblioteke koje se koriste su shinythemes [96] predlošci za estetski dio aplikacije, shiny web framework, DT [97] interface za data DataTable [43] javascript biblioteku, te zip za komprimiranje rezultata i izvornih datoteka [98].

U „ui" varijabli definira se korisničko sučelje koje se sastoji od naslova, poruka i raznih elemenata.

```
      ui <- fluidPage(
        titlePanel("Classification variable complexity parameter estimator"),
        theme = shinytheme("cerulean"),
        p("Tables larger then 1000x100 are not allowed to use on this server.
In case that you need using this software on larger tables, please use our
source code software and run t using Rscript."),
        p("To download data examples for testing purposes click here: " ,
            downloadLink("downloadDataExamples", "Download data examples")),
        p("Please use COMMA (,) separated CSV files. Columns which contain non
binary data will be skipped. Only digits 0 and 1 are allowed. Files need to
have column header!"),
```

Naslov aplikacije definira „titlePanel". Svojstvo „theme" koristi temu iz „shinytheme" biblioteke. Tekstualni dijelovi napisani su pomoću „p" naredbe.

Nakon naslova i teme, slijedi „sidebar" u kojem se nalazi objekt za izbor CSV datoteke i glavna ploča (mainPanel). CSV datoteka mora biti sastavljena od binarnih brojeva odvojeni zarezom.

```
        sidebarLayout(
          fileInput("file1"
                    , "Choose CSV File for analysis",
                    accept = c(
                      "text/csv", "text/comma-separated-values,text/plain",
".csv")
          ),
          mainPanel(
            downloadButton('downloadData', 'Download results'),
            p("All tabs below represent tables wich include descriptors and
complexity"),
              tabsetPanel(
                id = "tablePanel",
                tabPanel("Formula",
```

```
DT::dataTableOutput("columnComplexityFormulaData")),
                tabPanel("Simulation",
DT::dataTableOutput("columnComplexitySimulationData"))
              )
          )
```

Unutar glavne ploče nalazi se dugme za download podataka (downloadButton) kao i tekstualna poruka i tabovi u kojima se nalaze rezultati obrade koji se pojavljuju nakon unosa CSV podataka. Tabovi predstavljaju tablice rezultata dobivenu ili simulacijom ili formulom. Tablice se odnose na stupce onih podataka koji su dani u CSV formatu u aplikaciju i automatski obrađeni.

Sučelje služi samo za kontrolu uvoza podataka, pregled i njihov izvoz, a onaj dio koji to obrađuje u pozadini nalazi se u server funkciji.

(g) Definicija poslužitelja - DCAServer.R

Pozadinski rad aplikacije vrši se u DCAServer.R datoteci koja vrši dohvat, obradu i vraća rezultat sučelju.

```
server <- function(input, output, session) {
  dl=DocLibClass$new()
  dca = DCAClass$new()
  caRes = data.frame()
```

U „server" varijabli sprema se funkcija koja za argumente uzima sesiju te ulazne i izlazne varijable. U toj funkciji definiraju se pomoćne varijable i razne metode te izlazni podaci.

Jedna od metoda „runAnalysis" pokreće analizu ulaznih podataka, prebacuje ih u CSV format te ih komprimira pripremajući ih za izvođenje preko web sučelja.

```
 runAnalysis=function(fileName){
    caRes=dca$run(fileName = fileName)
    if(caRes$Error!=""){
      showNotification(caRes$Error
                       , duration = 10
                       , type = "error"
                       , closeButton = TRUE)
    }
    dca$export()
    dca$zipIt()
    return(caRes)
  }
```

Nakon komprimiranja „runAnalysis" metoda vraća rezultat obrade za potrebe prikaza.

Dohvat rezultata radi se preko „getData” metode.

```
getData=function(isRow,isSimulation,fileName){
```

Metoda „getData” provjerava jesu li rezultati dobiveni preko formule ili simulacijom te koji je naziv datoteke. Ukoliko datoteka s tim imenom postoji, učita je u memoriju i vrati rezultat. Međutim ako datoteka ne postoji, metoda „getData” napravi ponovnu obradu.

Kako bi podaci bili prikazani u obliku tablice, potrebno je podatke renderirati u oblik „DataTable“ čemu služe slijedeće linije. Ugnježđeni dijelovi koda su isključeni.

```
output$columnComplexityFormulaData = DT::renderDataTable({  ...
output$columnComplexitySimulationData = DT::renderDataTable({  ...
```

U slučaju da se podaci ne obnavljaju pravovremeno, svaki skup podataka koji treba biti prikazan i obnovljen, se „observe” funkcijom prati i unutar nje pozivaju metode za renderiranje podataka (DT::renderDataTable).

```
observe({
    if (is.null(input$file1$name)) return()

    if (!file.exists(getFileName()))  moveFileToCorrectName()
      cat("observe:",getFileName() , "\n")
...
```

Funkcija observer prilikom izvršavanja provjerava je li korisnik poslao datoteku, te ukoliko nije, prekine s radom. Ukoliko je datoteka dana od strane korisnika, premjesti je na odgovarajuće mjesto kako bi bilo lakše raditi s njom.

„Download source” link oslanja se na vrijednost „output$downloadSource” svojstva, čija se vrijednost dobiva pokretanjem „downloadHandler” funkcije koja nalazi sve .R datoteke i sve CSV datoteke iz direktorija u kojem je izvorni kod programa, komprimira ih ZIP algoritmom i sprema u „ca.zip” datoteku ukoliko ta datoteka već ne postoji. Ukoliko postoji, tada se tu datoteku prosljedi u fname varijablu i datoteka je prihvati kao sadržaj za izvoz.

```
  output$downloadSource = downloadHandler(
    filename = function() {   "ca.zip"  },
    content = function(fname) {
      if (!file.exists("ca.zip")){
        filesR=list.files(".","*.R")
        testFiles=list.files("test/","*.csv")
        testFiles=dl$filterArr(testFiles, "*row\\ *", toInclude = FALSE)
        testFiles=dl$filterArr(testFiles, "*col\\_*", toInclude = FALSE)
        testFiles=paste0("test/",testFiles)
```

```
              zip::zipr(zipfile = fname , files = c(testFiles,filesR) , recurse
= TRUE , include_directories = TRUE)
              return()
            }

            file.copy("ca.zip" ,fname)
          }
```

Drugi download odnosi se na zipane CSV datoteke dobivene obradom podataka.

```
        output$downloadData = downloadHandler(
```

Taj handler se pokreće klikom mišem na „Download results" dugme pri čemu se izvrši eksport podataka i šalju komprimirani podaci korisniku.

PRILOG 3.45 ProtSeqAnalizer

ProtSeqAnalizer predstavlja tehničko rješenje za obradu proteinskih sekvenci. Za ulazne proteinske/polipeptidne sekvence u CSV datoteci ProtSeqAnalizer vraća kao rezultat postojanje ili učestalost svih uzoraka unutar tih proteina. Izlazna datoteka može imati po desetke tisuća stupaca.

Izvorni kod aplikacije nalazi se na adresi http://meteo2.irb.hr/doktorat/ProtSeqAnalzyer.zip , [82].

Ulazna datoteka sadrži listu proteinskih sekvenci (primarne strukture) u kojoj sekvence nisu oblikovane u nekom od poznatih datotečnih formata nego svaka sekvenca zauzima jednu liniju bez obzira na njenu duljinu.

```
      Rscript ProtSeqAnalizerScript.R -m <method=simple> -i <inputFile> -b -l
<length> -p <preset=''> -c <customMask> -sz

      -b     Binarize - coverts any number >0 to 1

      -c     Custom mask. You can put mask like GX?G here where X and ?
represent any aminoacid but group them differently

      -l     length of motif (default length = 3)

      -i     input file - must be file with sequences stored as 1-line strings

      -o     Output file name

      -m     Method. Choose one of available methods: simple , masked,
custom_mask (default=simple)

      -p     Preset (default value '')

            presets: group hydrophobicity charge

      -sz    Skip rows which have only zero data.

      -s     Is secondary structure (default=0) - cannot work with presets
```

Aplikacija se pokreće pomoću Rscript naredbe i ProtSeqAnalizerScript.R skripte. Skripta uzima metodu '-m" argumentom, ulaznu datoteku "-i" argumentom, opciju kojom korisnik odabire hoće li izlaz biti predstavljen pomoću klasa 0 i 1 , ili kao integer "-b" argumentom, te duljinu motiva koji

pretražuje "'-l" argumentom . Odabir izlazne datoteke je automatski i generira koristeći trenutno vrijeme i datum ukoliko nije korišten „-o <imeIzlazneDatoteke>" kojem je potrebno napisati ime izlazne datoteke.

Aplikacija sadrži slijedeća računanja frekvencija motiva:

simple – prikaže učestalost svakog motiva na koji naiđe

masked – prikaže motive sa svim kombinacijama rupa označenih s "X"

custom_mask – korisnik sam stavlja motiv koji pretražuje npr. "GXXG" tada će dobiti kao rezultat 1 kolonu s učestalošću svih motiva koji odgovaraju „GXXG" motivu. Ukoliko umjesto „X" znaka, korisnik upiše „?" znak, tada će nastati više kolona nego u prethodnom slučaju pa će „?" predstavljati svaku nađenu aminokiselinu. Ukratko „X" će grupirati sve aminokiseline u jednu „X" kolonu , a „?" će ih svih prikazati zajedno s njihovim učestalostima

Opcija „-p" je opcionalna i odnosi se na grupiranje aminokiselina u razne grupe. Ta opcija je dostupna kod svih metoda, te radi na principu da aminokiseline određene grupe zamijeni rednim brojem te grupe pa takav uzorak pretražuje u sekvenci i prebrojava. U slučaju da redak ne sadrži niti jedan traženi motiv, tada je opcijom "-sz" moguće preskočiti taj redak.

Vrste podjela koje aplikacija koristi su: group , hydrophobicity i charge i pripadnost aminokiselinama tim skupovima se nalaze u „data/aaGroups.csv", gdje korisnik može mijenjati postojeće grupe ili dodavati nove.

Datoteka „aaGroups.csv" je pisana u CSV formatu, pa za svaku kolonu postoji značenje. U prvoj koloni je amino-kiselinski ostatak (jednoslovni kod), u drugoj je troslovni kod aminokiseline, a ostale kolone odnose se na razna grupiranja aminokiselina. Korisnik može grupirati aminokiseline na način koji mu odgovara tako da mijenja postojeće grupe ili doda neku svoju kao novu kolonu.

PRILOG 3.46 Tablica za rangiranje modela prema finalnoj fazi (IS3) s izazova za predviđanje tumora (Tumour prediction challenge [80]) prema $F1$ parametru iz [79], $Q2$, MCC, and $\Delta Q2$ [35].

Tablica sadrži slijedeće stupce:

No. - redni broj retka

ID – identifikator modela

TP – broj točnih predviđanja klase 1

TN – broj točnih predviđanja klase 0

FP– broj slučajeva kad je klasa 0 predviđena kao klasa 1

FN – broj slučajeva kada je klasa 1 predviđena kao klasa 0

F1score – vrijednost parametra *F1*

Q2 – vrijednost Q2 parametra

MCC – vrijednost Matthewsovog koeficijenta korelacije (*MCC*)

ΔQ2 – vrijednost doprinosa modela

No.	ID	TP = *p*	TN = *n*	FP = *o*	FN = *u*	F1 = F1score	Q2	*MCC*	ΔQ2
1	X2463247	7335	16506	278	568	94,55	96,57	0,921	39,68
2	X2478107	7308	16561	223	595	94,7	96,69	0,923	39,674
3	X2453885	7291	16594	190	612	94,79	96,75	0,925	39,666
4	X2478109	7274	16613	171	629	94,79	96,76	0,925	39,621
5	X2473029	7253	16643	141	650	94,83	96,8	0,926	39,583
6	X2472860	7255	16637	147	648	94,81	96,78	0,926	39,579
7	X2463211	7253	16622	162	650	94,7	96,71	0,924	39,529
8	X2456287	7226	16666	118	677	94,79	96,78	0,926	39,494
9	X2456202	7202	16689	95	701	94,76	96,78	0,926	39,422
10	X2476556	7191	16711	73	712	94,82	96,82	0,927	39,418
11	X2468117	7203	16677	107	700	94,7	96,73	0,925	39,396
12	X2470044	7192	16700	84	711	94,76	96,78	0,926	39,395
13	X2460633	7207	16657	127	696	94,6	96,67	0,923	39,366
14	X2476415	7178	16705	79	725	94,7	96,74	0,925	39,331
15	X2476341	7169	16722	62	734	94,74	96,78	0,926	39,326
16	X2467165	7196	16649	135	707	94,47	96,59	0,921	39,285
17	X2476776	7156	16714	70	747	94,6	96,69	0,924	39,233
18	X2460162	7149	16710	74	754	94,53	96,65	0,923	39,184
19	X2466534	7212	16567	217	691	94,08	96,32	0,915	39,16
20	X2453883	7140	16695	89	763	94,37	96,55	0,921	39,096
21	X2470029	7132	16709	75	771	94,4	96,57	0,921	39,088
22	X2467034	7156	16640	144	747	94,14	96,39	0,917	39,041
23	X2470090	7198	16537	247	705	93,8	96,14	0,911	39,006
24	X2465312	7150	16619	165	753	93,97	96,28	0,914	38,954
25	X2469655	7149	16606	178	754	93,88	96,22	0,913	38,915
26	X2420793	7061	16741	43	842	94,1	96,42	0,918	38,78
27	X2453909	7099	16636	148	804	93,72	96,14	0,911	38,717
28	X2450373	7250	16311	473	653	92,79	95,44	0,895	38,706
29	X2450371	7075	16588	196	828	93,25	95,85	0,904	38,46
30	X2463232	7439	15798	986	464	91,12	94,13	0,868	38,416
31	X2445210	7005	16703	81	898	93,47	96,03	0,909	38,373
32	X2446437	7005	16703	81	898	93,47	96,03	0,909	38,373
33	X2453907	7005	16703	81	898	93,47	96,03	0,909	38,373
34	X2456285	6995	16721	63	908	93,51	96,07	0,91	38,365
35	X2426377	7190	16133	651	713	91,34	94,47	0,873	37,914
36	X2420789	7186	16133	651	717	91,31	94,46	0,872	37,892
37	X2436976	7093	16304	480	810	91,66	94,77	0,879	37,823
38	X2431851	7010	16438	346	893	91,88	94,98	0,884	37,713
39	X2426321	7277	15790	994	626	89,98	93,44	0,852	37,503
40	X2455084	6874	16622	162	1029	92,03	95,18	0,889	37,441
41	X2449602	6887	16586	198	1016	91,9	95,08	0,887	37,42
42	X2431183	7034	16185	599	869	90,55	94,05	0,862	37,189
43	X2430956	7344	15498	1286	559	88,84	92,53	0,834	37,115
44	X2427522	6934	16345	439	969	90,78	94,3	0,868	37,053
45	X2445117	6794	16628	156	1109	91,48	94,88	0,882	37,016
46	X2430887	6674	16699	85	1229	91,04	94,68	0,878	36,54
47	X2427156	6975	16046	738	928	89,33	93,25	0,844	36,504
48	X2453859	6633	16758	26	1270	91,1	94,75	0,881	36,467
49	X2425329	6980	15834	950	923	88,17	92,41	0,826	35,982
50	X2453761	6560	16475	309	1343	88,82	93,31	0,845	35,331
51	X2470450	6489	16541	243	1414	88,68	93,29	0,845	35,111
52	X2875297	6489	16423	361	1414	87,97	92,81	0,833	34,805
53	X2449297	6328	16589	195	1575	87,73	92,83	0,835	34,348
54	X2875291	7359	14257	2527	544	82,74	87,56	0,743	33,979
55	X2453752	6204	16676	108	1699	87,29	92,68	0,833	33,891
56	X2875294	6389	16263	521	1514	86,26	91,76	0,808	33,839
57	X2453861	6129	16765	19	1774	87,24	92,74	0,835	33,709
58	X2464413	6205	16600	184	1698	86,83	92,38	0,825	33,7
59	X2453737	6117	16726	58	1786	86,9	92,53	0,83	33,542
60	X2464535	6152	16621	163	1751	86,54	92,25	0,822	33,462
61	X2445581	6212	16161	623	1691	84,3	90,63	0,781	32,6
62	X2467686	6176	16219	565	1727	84,35	90,72	0,783	32,552
63	X2463868	5686	16730	54	2217	83,35	90,8	0,791	31,178
64	X2460591	6607	14355	2429	1296	78,01	84,91	0,669	30,091
65	X2463397	5721	16056	728	2182	79,72	88,21	0,723	29,623
66	X2463748	5350	16099	685	2553	76,77	86,88	0,691	27,691
67	X2470438	4969	16775	9	2934	77,15	88,08	0,73	27,346
68	X2405012	5023	16468	316	2880	75,86	87,05	0,699	26,847
69	X2410591	6798	12296	4488	1105	70,85	77,34	0,555	25,803
70	X2469987	5909	13547	3237	1994	69,32	78,81	0,536	24,151

PRILOG 3.47 Tablica rangiranje modela prema finalnoj fazi (IS3) s izazova za predviđanje tumora (Tumour prediction challenge [99]) prema $F1score$ parametru iz [20], $Q2, MCC$, i $\Delta Q2$

Tablica je nastavak prethodne tablice u *Prilogu 3.46*. U ovoj tablici prikazani su slijedeći parametri:

No. - redni broj retka

ID – identifikator modela

rank_F1score – redosljed prema F1 parametru

rank_Q2b – redosljed prem Q2 parametru

rank_*MCC* – redosljed prem *MCC* parametru

rank_ΔQ2 – redosljed prem ΔQ2 parametru

diff-rank_F1score – razlika u rangiranju u odnosu na ΔQ2 parametar

diff-rank_Q2b – razlika u rangiranju u odnosu na ΔQ2 parametar

diff-rank_*MCC* – razlika u rangiranju u odnosu na ΔQ2 parametar

No.	***ID***	rank_F1score	rank_Q2b	rank_*MCC*	rank_ΔQ2	diff-rank_F1score	diff-rank_Q2b	diff-rank_*MCC*
1	X2463247	16	19	19	**1**	15	18	18
2	X2478107	10	14	14	**2**	8	12	12
3	X2453885	**5**	9	10	**3**	2	6	7
4	X2478109	**4**	8	9	**4**	0	4	5
5	X2473029	**1**	**2**	**2**	**5**	4	3	3
6	X2472860	**3**	***4***	7	6	3	2	1
7	X2463211	11	12	12	7	4	5	5
8	X2456287	6	***4***	6	8	2	4	2
9	X2456202	7	*6,5*	**5**	9	2	2,5	4
10	X2476556	**2**	**1**	**1**	10	8	9	9
11	X2468117	13	11	11	11	2	0	0
12	X2470044	8	***4***	**4**	12	4	8	8
13	X2460633	15	15	15	13	2	2	2
14	X2476415	12	10	8	14	2	4	6
15	X2476341	9	*6,5*	**3**	15	6	8,5	12
16	X2467165	18	17	17	16	2	1	1
17	X2476776	14	13	13	17	3	4	4
18	X2460162	17	16	16	18	1	2	2
19	X2466534	23	23	23	19	4	4	4
20	X2453883	20	20	20	20	0	0	0
21	X2470029	19	18	18	21	2	3	3
22	X2467034	21	22	22	22	1	0	0
23	X2470090	26	27	27	23	3	4	4
24	X2465312	24	24	24	24	0	0	0
25	X2469655	25	25	25	25	0	0	0
26	X2420793	22	21	21	26	4	5	5
27	X2453909	27	26	26	27	0	1	1
28	X2450373	33	33	33	28	5	5	5
29	X2450371	32	32	32	29	3	3	3
30	X2463232	41	44	43	30	11	14	13
31	X2445210	29	29	29	31	2	2	2
32	X2446437	30	30	30	32	2	2	2
33	X2453907	31	31	31	33	2	2	2
34	X2456285	28	28	28	34	6	6	6
35	X2426377	39	41	41	35	4	6	6
36	X2420789	40	42	42	36	4	6	6
37	X2436976	37	38	39	37	0	1	2
38	X2431851	36	36	36	38	2	2	2
39	X2426321	46	46	46	39	7	7	7
40	X2455084	34	34	34	40	6	6	6
41	X2449602	35	35	35	41	6	6	6
42	X2431183	45	45	45	42	3	3	3

43	X2430956	48	55	52	43	5	12	9
44	X2427522	44	43	44	44	0	1	0
45	X2445117	38	37	37	45	7	8	8
46	X2430887	43	40	40	46	3	6	6
47	X2427156	47	49	49	47	0	2	2
48	X2453859	42	39	38	48	6	9	10
49	X2425329	51	56	56	49	2	7	7
50	X2453761	49	47	48	50	1	3	2
51	X2470450	50	48	47	51	1	3	4
52	X2875297	52	51	53	52	0	1	1
53	X2449297	53	50	51	53	0	3	2
54	X2875291	63	65	63	54	9	11	9
55	X2453752	54	53	54	55	1	2	1
56	X2875294	59	59	59	56	3	3	3
57	X2453861	55	52	50	57	2	5	7
58	X2464413	57	57	57	58	1	1	1
59	X2453737	56	54	55	59	3	5	4
60	X2464535	58	58	58	60	2	2	2
61	X2445581	61	62	62	61	0	1	1
62	X2467686	60	61	61	62	2	1	1
63	X2463868	62	60	60	63	1	3	3
64	X2460591	65	68	68	64	1	4	4
65	X2463397	64	63	65	65	1	2	0
66	X2463748	67	67	67	66	1	1	1
67	X2470438	66	64	64	67	1	3	3
68	X2405012	68	66	66	68	0	2	2
69	X2410591	69	70	69	69	0	1	0
70	X2469987	70	69	70	70	0	1	0

PRILOG E_3 (Elektronički prilozi)

U ovom dijelu nalaze se poveznice na adrese aplikacija koje su izrađene u svrhu ovog rada.

PRILOG E_3.1 *Classification variable complexity parameter estimator*

Aplikacija je objašnjena u *Prilogu 3.44.*
Nalazi se na adresi: http://meteo2.irb.hr/shiny/CA/
Izvorni kod je dostupan na adresi:
http://meteo2.irb.hr/doktorat/E_3.1(complexityEstimator.Final).zip

---

PRILOG E_3.2 *Zagreb indices and their modifications – CALCULATOR – Web application*

Aplikacija je objašnjena u (*Prilogu 3.43*).
Adresa izvornog koda web aplikacije je: http://meteo2.irb.hr/doktorat/E_3.2_ZagrebIndicesWeb.zip
U bin folderu aplikacije, nalazi se binarna verzija programa, te ju je moguće pokrenuti iz komandne linije neovisno o operativnom sustavu.
Adresa izvornog koda java konzolne aplikacije je: https://github.com/vbojovic1980/sci-rep i nalazi se u folderu randIndex.
Adresa na kojoj je aplikacija dostupna za korištenje kao mrežni poslužitelj je http://meteo2.irb.hr/indexer/.

PRILOG E_3.3 *ProtSeqAnalizer*

Aplikacija je objašnjena u (*Prilogu 3.45*). Izvorni kod aplikacije se nalazi na adresi: http://meteo2.irb.hr/doktorat/E_3.3_ProtSeqAnalzyer.zip

---

PRILOG E_3.4 Simulator

Aplikacija je objašnjena u (*Prilogu 3.41*). Izvorni kod i aplikacija , te podaci simulacija se nalaze na adresi: http://meteo2.irb.hr/doktorat/E_3.4_scirep.zip

---

PRILOG E_3.5 Jednostavni Simulator

Aplikacija je objašnjena u (*Prilogu 3.42*). Izvorni kod i aplikacija , te podaci simulacija se nalaze na adresi : http://meteo2.irb.hr/doktorat/E_3.5_scirep.zip u folderu HSKIKI.

PRILOG E_3.6 Tablice iz rada „Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology“ [66]

U ovom prilogu nalaze se dvije tablice, jedna sa 58 deskriptora, a druga s 30 deskriptora. Nalaze se na adresi:

- http://meteo2.irb.hr/doktorat/E_3.6_Huuskonen30.zip
- http://meteo2.irb.hr/doktorat/E_3.6_Huuskonen58.zip

PRILOG E_3.7 Tablice ovisnosti normiranih parametara $logW$ i $\Delta Q_{2,max}$.o N i x

Konstruirani podaci s vrijednostima udjela klase 1 x i normiranih parametara $logW$ i $\Delta Q_{2,max}$.u ovisnosti o $N \in [15,2000]$ nalaze se na adresi :
http://meteo2.irb.hr/doktorat/E_3.7_(postociAll).csv - tablica s udjelima i normiranim parametrima bez grupiranja. Stupci u tablici su slijedeći: N– broj podataka u varijabli, x – udio klase „1“, $logW$ – prilagođena formula entropije, nx – broj jedinica u varijabli
http://meteo2.irb.hr/doktorat/E_3.7_(analiza_joinedGrp).csv - grupirana tablica s udjelima koja sadrži minimalne i maksimalne vrijednosti x – a za svaki od nivoa vrijednosti (0.01, 0.025, 0.05, 0.075 i 0.1). Stupci u tablicu su slijedeći: N – broj podataka u varijabli, $level$ – nivo vrijednosti normiranih parametara ($logW$ ili $\Delta Q_{2,max}$), min_xn – najmanji broj podataka klase 1 potreban da se postigne odgovarajući nivo, min_x – najmanji udio klase 1 potreban da se postigne odgovarajući nivo, max_xn – najveći udio klase 1 potreban da se postigne odgovarajući nivo, max_x– najveći broj podataka klase 1 potreban da se postigne odgovarajući nivo, $varName$ – naziv parametra ($logW$ ili $\Delta Q_{2,max}$), comparison – oznaka koji je parametar veći ($logW$ ili $\Delta Q_{2,max}$).

PRILOG E_3.8 Tablice deskriptora DADP (*engl*. Database of Anuran Defense Peptides) baze [65]

U ovom prilogu nalazi se nekoliko tablica komprimirano u jednu datoteku.

- http://meteo2.irb.hr/doktorat/E_3.8_velikaTablica.zip

U datoteci se nalaze csv datoteke zajedno sa izračunima složenosti dobivene pomoću estimatora (nastavak .col_formula.csv i .col_simulation.csv) koje su dobivene ili formulama za računanje karakterističnih vrijednosti ili simulacijom, tj. različitim poretkom varijabli. [76] Znakom X označeno je grupiranje svake aminokiseline, a znakom ? – bilokoja aminokiselina negrupirana. Motivi su dobiveni pomoću programa ProtSeqAnalizer (*Prilog E_3.3*) [82], koristeći sekvence iz DADP baze [65], a valencije veza su računane pomoću kalkulatora indeksa [53].

- aa2.csv– broj motiva veličine 2
- aa.csv – broj motiva veličine 2 grupirani neovisno o poretku aminokiselina
- amino.csv – broj pojedine aminokiseline u polipeptidu
- axxa.csv – broj motiva koji se dobivaju izrazom ?XX?, tj. motivi od AXXA do ZXXZ
- axxxa.csv – broj motiva koji se dobivaju izrazom ?XXX?, tj. motivi od AXXXA do ZXXXZ
- desc.csv – različiti deskriptori
- GXXG.csv – broj GXXG motiva
- kontakti.csv – valencije veza od
- prve.csv – podaci preuzeti sa DADP stranice [65]
- sheet1.csv – sve nabrojane datoteke u jednoj
- GXXXG.txt – broj GXXXG motiva

11. ŽIVOTOPIS I POPIS PUBLIKACIJA

Viktor Bojović rođen je 1980. u Splitu gdje završava srednju elektrotehničku školu s maturalnom temom računalnih virusa. Studij profesora informatike i tehničke kulture završava 2007. godine na Fakultetu prirodoslovno-matematičkih znanosti i kineziologije Sveučilišta u Splitu s temom „Prilagodba PDB baze relacijskom modelu radi lakše vizualizacije membranskih proteina“ (mentor: prof. dr. sc. Marko Rosić, komentor: prof. dr. sc. Davor Juretić). Nakon toga kratko radi na Institutu Ruđer Bošković u Zagrebu kao stručni suradnik u Centru za informatiku i računarstvo, ali nastavlja suradnju i dalje kao vanjski suradnik IRB-a (dr. sc K. Skala, dr. sc. B. Lučić, dr. sc. N. Trinajstić) i splitskog PMF-a (prof. dr. sc. D. Juretića, dr. sc. I. Ljubenkov), dok za to vrijeme radi uglavnom u realnom sektoru kao programer na razvoju raznoga softvera i baza podataka, a povremeno i kao nastavnik na zamjenama u srednjim školama. Od listopada 2018. zaposlen je kao doktorand na Institutu Ruđer Bošković uz potporu Europske unije putem Europskog socijalnog fonda i Republike Hrvatske putem Hrvatske zaklade za znanost. Posjeduje aktivna znanja u programiranju u raznim programskim jezicima, alatima, i okruženjima. Ta iskustva pomogla su mu da, radom kao vanjski suradnik kroz suradnje s više istraživačkih grupa, aktivno sudjeluje u istraživanjima iz područja bioinformatike i kemoinformatike u modeliranju strukture i svojstava proteina, peptida i organskih kemijskih spojeva. Su-autor je šest radova u časopisima koje navodi baza *Current Contents*, četiri rada u zbornicima znanstvenih skupova koje navodi baza *Web of Science* (kao prvi autor), jednog poglavlja u knjizi, i 11 sažetaka u zbornicima znanstvenih skupova. Osobno je sudjelovao je na četiri znanstvena skupa s predavanjima na engleskom jeziku, te na jednom međunarodnom znanstvenom skupu s posterom.

Poglavlja u knjigama

1. Juretić, Davor; Tossi, Alessandro; Kamech, Nédia; Ilić, Nada; Bojović, Viktor; Novković, Mario; Simunić, Juraj; Petrov, Dražen; Lučić, Bono; Miljak, Marija et al.
From Data Collecting to Web servers for Automatic Design of Peptide Antibiotics. //
Bioinformatics and biological physics : proceedings of the scientific meeting / Paar, Vladimir (ur.).
Zagreb: Hrvatska akademija znanosti i umjetnosti, 2013. str. 63-78. (https://www.bib.irb.hr/680096)

Znanstveni i pregledni radovi u časopsima koje navodi baza *Current Contents*

1. Lučić, Bono; Batista, Jadranko; Bojović, Viktor; Lovrić, Mario; Sović Kržić, Ana; Bešlo, Drago; Nadramija, Damir; Vikić-Topić, Dražen
Estimation of random accuracy and its use in validation of predictive quality of classification models within predictive challenges. // Croatica chemica acta, 92 (2019), 3; 379-391
doi:10.5562/cca3551 (međunarodna recenzija, članak, znanstveni), https://hrcak.srce.hr/238284

2. Bilušić, Tea; Žanetić, Mirella; Ljubenkov, Ivica; Generalić Mekinić, Ivana; Štambuk, Snježana; Bojović, Viktor; Soldo, Barbara; Magiatis, Prokopios
Molecular characterization of Dalmatian cultivars and the influence of the olive fruit harvest period

on chemical profile, sensory characteristics and oil oxidative stability. // European food research and technology, 244 (2018), 2; 281-289 doi:10.1007/s00217-017-2954-7 (međunarodna recenzija, članak, znanstveni)

3. Bošnjak, Ivana; Bojović, Viktor; Šegvić-Bubić, Tanja; Bielen, Ana
Occurrence of protein disulfide bonds in different domains of life : a comparison of proteins from the Protein Data Bank. // Protein engineering, design & selection, 27 (2014), 3; 65-72 doi:10.1093/protein/gzt063 (međunarodna recenzija, članak, znanstveni)

4. Novković, Mario; Simunić, Juraj; Bojović, Viktor; Tossi, Alessandro; Juretić, Davor
DADP: the database of anuran defense peptides. // Bioinformatics, 28 (2012), 10; 1406-1407 doi:10.1093/bioinformatics/bts141 (međunarodna recenzija, članak, znanstveni)

5. Kamech, Nédia; Vukičević, Damir; Ladram, Ali; Piesse, Christophe; Vasseur, Julie; Bojović, Viktor; Simunić, Juraj; Juretić, Davor
Improving the Selectivity of Antimicrobial Peptides from Anuran Skin. // Journal of chemical information and modeling, 52 (2012), 12; 3341-3351 doi:10.1021/ci300328y (međunarodna recenzija, članak, znanstveni)

6. Juretić, Davor; Vukičević, Damir; Petrov, Dražen; Novković, Mario; Bojović, Viktor; Lučić, Bono; Ilić, Nada; Tossi, Alessandro
Knowledge-based computational methods for identifying or designing novel, non-homologous antimicrobial peptides. // European biophysics journal, 40 (2011), 4; 371-385 doi:10.1007/s00249-011-0674-7 (međunarodna recenzija, pregledni rad, znanstveni)

Znanstveni radovi objavljeni u zbornicima skupova

1. Bojović, Viktor; Lučić, Bono; Bešlo, Drago; Skala, Karolj; Trinajstić, Nenad
Calculation of topological molecular descriptors based on degrees of vertices. // 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)
Opatija, Hrvatska, 2019. str. 266-269 doi:10.23919/MIPRO.2019.8757128 (predavanje održao V. Bojović, međunarodna recenzija, cjeloviti rad (in extenso), znanstveni)
https://ieeexplore.ieee.org/document/8757128

2. Bojović, Viktor; Sović, Ivan; Bačić, Andrea; Lučić, Bono; Skala, Karolj
A novel tool/method for visualization of orientations of side chains relative to the protein's main chain. // Proceedings Vol. I. MEET&GVS 34rd International Convention MIPRO 2011 / Biljanović, Petar ; Skala, Karolj ; (ur.).
Zagreb: Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, 2011. str. 273-276 (predavanje održao V. Bojović, međunarodna recenzija, cjeloviti rad (in extenso), znanstveni)

3. Bojović, Viktor; Lučić, Bono; Skala, Karolj; Grubišić, Ivan
Analyser and viewer of protein inter-residue contacts. // MIPRO 2010, MEET&GVS / Biljanović, Petar ; Skala, Karolj ; (ur.).
Zagreb: DENONA, 2010. str. 299-302. (https://www.bib.irb.hr/489666) (predavanje održao V. Bojović, međunarodna recenzija, cjeloviti rad (in extenso), znanstveni)

4. Bojović, Viktor; Lučić, Bono; Skala, Karolj
Protein Data Bank Graphics Generator on Grid. // Proceedings Vol. I. MEET&GVS 32rd International Convention MIPRO 2009 / Biljanović, Petar ; Skala, Karolj (ur.).
Rijeka: Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, 2009. str. 341-345 (predavanje održao V. Bojović, međunarodna recenzija, cjeloviti rad (in extenso), znanstveni)

Sažeci u zbornicima skupova

1. Lučić, Bono; Batista, Jadranko; Bojović, Viktor; Lovrić, Mario
Novel statistical parameters for model quality estimation. // Book of abstract, Math/Chem/Comp 2019 / Vančik, Hrvoje ; Cioslowski, Jerzy (ur.).
Zagreb, 2019. str. 2-2 (pozvano predavanje, međunarodna recenzija, sažetak, znanstveni)

2. Bojović, Viktor; Batista, Jadranko; Amić, Ana; Lučić, Bono
Estimation of the random correlation level of molecular descriptors in structure- property modeling. // Knjiga sažetaka/Book of Abstracts / Galić, Nives ; Rogošić, Marko (ur.).
Zagreb: Hrvatsko društvo kemijskih inženjera i tehnologa, 2019. str. 153-153 (poster, domaća recenzija, sažetak, znanstveni)

3. Bielen, Ana; Bojović, Viktor; Šegvić-Bubić, Tanja; Bošnjak, Ivana
Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank. // Book of abstracts of FEBS EMBO 2014 Conference
Pariz, Francuska, 2014. (poster, međunarodna recenzija, sažetak, znanstveni)

4. Novković, Mario; Simunić, Juraj; Bojović, Viktor; Tossi, Alessandro; Juretić, Davor
DADP: the Database of Anuran Defense Peptides. // New Antimicrobials Workshop in Trieste, 25-26.5.2012 / Tossi, Alessandro (ur.).
Trieste: Edizioni Universita di Trieste, 2012. str. 39-40 (poster, međunarodna recenzija, sažetak, znanstveni)

5. Juretić, Davor; Vukičević, Damir; Ilić, Nada; Novković, Mario; Simunić, Juraj; Bojović, Viktor; Kamech, Nédia; Tossi, Alessandro
In silico search strategies for finding new AMPs. // New Antimicrobials Workshop in Trieste, 25-26.5.2012 / Tossi, Alessandro (ur.).
Trst: Edizioni Universita di Trieste, 2012. str. 23-24 (pozvano predavanje, međunarodna recenzija, sažetak, znanstveni)

6. Lučić, Bono; Bojović, Viktor; Vukičević, Damir; Juretić, Davor
Positional specific amino acid attributes improve the computational design of antimicrobial peptides. // Australia-Croatia Workshop on Antimicrobial Peptides and Summer School in Biophysics in Biophysics (AMP2010) : Book of Abstracts / Juretić, Davor ; Šeparović, Frances (ur.).
Split: Faculty of Sciences, 2010. str. 43-43 (poster, sažetak, znanstveni)

7. Juretić, Davor; Vukičević, Damir; Ilić, Nada; Bojović, Viktor; Tossi, Alessandro
Design optimisation of novel peptide antibiotics predicted to have high selectivity against G-negative bacteria. // The 3rd Adriatic Meeting on Computational Solutions in the Life Sciences / Tomić, Sanja ; Smith, David (ur.).
Zagreb: Centre for Computational Solutions in the Life Sciences, Ruđer Bošković Institute, 2009. str. 59-59 (poster, sažetak, znanstveni)

8. Juretić, Davor; Vukičević, Damir; Bojović, Viktor; Lučić, Bono; Ilić, Nada
Design principles for peptide antibiotics. // From Solid State to Biophysics, 4th conference, Cavtat, Croatia, 2008
Cavtat, Hrvatska, 2008. str. x-x (poster, sažetak, znanstveni)

9. Juretić, Davor; Bojović, Viktor; Lučić, Bono
Sequence attributes for estimating the therapeutic index of peptide antibiotics. // The 2nd Opatija Meeting on Computational Solutions in the Life Sciences / Darko Babić, Nađa Došlić, David Smith, Sanja Tomić, Kristian Vlahoviček (ur.).
Zagreb: Ruđer Bošković Institute, Zagreb, Croatia, 2007. str. 19-19 (pozvano predavanje, međunarodna recenzija, sažetak, znanstveni)

10. Juretić, Davor; Lučić, Bono; Bojović, Viktor; Ilić, Nada
Designing peptide antibiotics with potential medical applications. // The 1st Split Meeting on Development and Applications of Novel Methods and Models in Computational Biophysics and Structural Bioinformatics
Split, Hrvatska, 2007. (pozvano predavanje, međunarodna recenzija, sažetak, znanstveni)

11. Juretić, Davor; Bojović, Viktor; Lučić, Bono; Ilić, Nada
Predicting the therapeutic index of peptide antibiotics. // Third Austrian-Croatian Science Days, Grac, Austrija, listopad 2007
Grac, Austrija, 2007. (poster, sažetak, znanstveni)