Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Domen Rački

Ocenjevanje atributov v neuravnoteženih problemih

diplomsko delo

visokošolski strokovni študijski program prve stopnje računalništvo in informatika

Mentor: prof. dr. Marko Robnik-Šikonja

Ljubljana, 2011

Rezultati diplomskega dela so intelektualna lastnina avtorja in Fakultete za računalništvo in informatiko Univerze v Ljubljani. Za objavljanje ali izkoriščanje rezultatov diplomskega dela je potrebno pisno soglasje avtorja, Fakultete za računalništvo in informatiko Univerze v Ljubljani ter mentorja.

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 00133/2011
Datum: 01.09.2011

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **DOMEN RAČKI**

Naslov: **OCENJEVANJE ATRIBUTOV V NEURAVNOTEŽENIH PROBLEMIH**
ATTRIBUTE EVALUATION ON IMBALANCED DATA SETS

Vrsta naloge: Diplomsko delo visokošolskega strokovnega študija prve stopnje

Tematika naloge:

Neuravnoteženost razreda je vir mnogih težav v strojnem učenju. Pojavlja se v številnih pomembnih realnih problemih; na primer, pri zavarovalniških goljufijah je v velikanski množici transakcij le nekaj primerov goljufij, pri vrednotenju tehničnih sistemov je napačnih delovanj le malo, itd. Za ocenjevanje atributov v tovrstnih primerih je bilo razvitih nekaj novih metod, ki v cenitveno funkcijo vsilijo predpostavko uniformne porazdelitve razredov. Na ta način metode bolj smiselno ocenjujejo atribute v neuravnoteženih problemih in so potencialno uspešnejše od obstoječih metod.
Testirajte novo razvite metode v odločitvenih drevesih in naključnih gozdovih na več množicah podatkov in pri različnih stopnjah neuravnoteženosti. Primerjajte AUC naučenih klasifikatorjev z obstoječimi metodami ocenjevanja atributov ter razlike statistično ovrednotite in jih grafično predstavite.

Mentor:
prof. dr. Marko Robnik Šikonja

Dekan:
prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Domen Rački, z vpisno številko **63080429**, sem avtor diplomskega dela z naslovom:

Ocenjevanje atributov v neuravnoteženih problemih

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Marka Robnik-Šikonje;
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela;
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 15. septembra 2011 Podpis avtorja:

Iskreno se zahvaljujem mentorju prof. dr. Marku Robnik-Šikonji za vso pomoč, izkazano potrpljenje in strokovno vodenje pri izdelavi diplomskega dela.

Hvala mami Majdi, ki ji nikoli noben izziv ni bil prevelik, in očetu Jožetu, ki mu nikoli nobena pot ni bila predolga, ker sta vedno verjela vame.

Hvala tudi vsem drugim, ki jih v zahvali nisem izrecno omenil za vso izkazano ljubezen in podporo v času študija.

Majdi in Jožetu.

Kazalo

Povzetek

V diplomskem delu smo analizirali delovanje mer za ocenjevanje atributov pri različnih stopnjah neuravnoteženosti porazdelitve razredov. Realne domene smo vzorčili tako, da smo dobili neuravnotežene dvorazredne množice v razmerjih 1:5, 1:10, 1:50, 1:100, 1:500 in 1:1000. Na teh množicah smo z izbranimi merami za ocenjevanje atributov zgradili modele odločitvenih dreves in s sorazmernim prečnim preverjanjem reda 5×2 izmerili AUC. Da bi ugotovili ali se uporabljene mere med seboj razlikujejo, smo na dobljenih rezultatih uporabili Friedmanov test, z Nemenyievim testom pa smo določili in grafično prikazali, med katerimi uporabljenimi merami obstajajo značilne razlike in katere mere delujejo podobno.

Kot najboljša mera pri naravnem razmerju razredov se izkaže MDL, za razmerje 1:5 pa kotna razdalja. Pri razmerjih 1:10 in 1:50 se je najboljše izkazala mera ReliefF, pri razmerjih 1:100, 1:500 in 1:1000 pa informacijski prispevek. Pri vseh razmerjih se kot najslabša mera izkaže klasifikacijska točnost.

Ključne besede:

strojno učenje, neuravnotežene množice, ocenjevanje atributov, CORElearn, odločitvena drevesa

Abstract

We analyze the performance of attribute evaluation measures on imbalanced datasets at different levels of imbalance. We sample real world datasets at ratios 1:5, 1:10, 1:50, 1:100, 1:500 and 1:1000. We build decision tree models and for each attribute evaluation measure compute AUC with stratified 5×2 cross validation. To test significance of the difference we use Friedman's test. With Nemenyi's test we determine and graphically display the similarities and differences.

We find that the best performing measure at unaltered class ratios is MDL, for class ratios 1:5 the best measure is the angular distance. For ratios 1:10 and 1:50 the beast measure is ReliefF and for class ratios 1:100, 1:500 and 1:1000 the best performing measure is information gain. The worst performing measure on all class ratios is accuracy.

Keywords:

machine learning, imbalanced datasets, attribute evaluation, CORElearn, decision trees

Poglavje 1

Uvod

Problem neuravnoteženih množic lahko ilustriramo z goljufi pri zavarovalniških poslih. Teh je v veliki množici zavarovancev precej manj kot poštenjakov. Interes zavarovalnice za odkrivanje goljufov je velik. Toda katera je tista lastnost, po kateri lahko zavarovalnica loči goljufa od poštenjakov?

Množici, v kateri imamo veliko primerov enega razreda — v našem primeru poštenjakov —, in malo primerov drugega razreda — v našem primeru goljufov —, pravimo neuravnotežena množica. Pri slednjih je klasifikacija lahko težavna, saj večinski razred ob izbiri neustrezne mere za ocenjevanje atributov premočno vpliva na gradnjo odločitvenega modela in s tem povzroči slabše ločevanje manjšinskega razreda. V našem primeru bi to pomenilo, da bi, ob neustrezni izbiri mere za ocenjevanje atributov, skozi mrežo smuknilo kar nekaj goljufov, ki bi jih smatrali za poštenjake. Ta problem rešujemo z uporabo mer za ocenjevanje atributov, prilagojenih neuravnoteženim množicam. V diplomskem delu smo analizirali in primerjali mere za ocenjevanje atributov na neuravnoteženih množicah.

Pripravili smo učne množice iz UCI ML Repository, LIBSVM in Delve repozitorijev in jih vzorčili tako, da smo dobili različna razmerja med večinskim in manjšinskim razredom. Z izbranimi merami za ocenjevanje atributov smo zgradili odločitvena drevesa in s sorazmernim prečnim preverjanjem reda 5×2 izmerili AUC. Dobljene ocene AUC modelov smo med seboj primer-

jali z Friedmanovim testom, podrobnejše razlike ugotavljali z Nemenyievmi testom in na koncu rezultate grafično predstavili. Pri delu smo uporabljali program Microsoft Excel za pripravo domen, v statističnemu programu R s paketom CORElearn smo zgradili in ovrednotili modele odločitvenih dreves in naključnih gozdov ter izračunali Friedmanovo statistiko. Ključne razlike po Nemenyievem testu smo prikazali s pomočjo programa Orange in programskega jezika Python.

V drugem poglavju diplomskega dela opisujemo uporabljene mere za ocenjevanje atributov. V tretjem poglavju sledi podrobnejši opis metodologije testiranja in pregled uporabljenih orodji. V četrtem poglavju so prikazani rezultati diplomskega dela. Na koncu podamo sklepne ugotovitve in ideje za izboljšave.

Poglavje 2

Mere za ocenjevanje atributov

Mere za ocenjevanje atributov izbirajo atribute za gradnjo odločitvenega drevesa glede na oceno kvalitete atributa. Čim večja je slednja, tem bolje atribut loči razrede.

Za ocenjevanje atributov in gradnjo odločitvenih modelov smo uporabili mere: ReliefF, MDL; klasične mere: informacijski prispevek, Gini-indeks, klasifikacijsko točnost in DKM, katerih opis je povzet po [6], [9] in [4]. Izmed mer z enakomerno apriorno porazdelitvijo smo preizkusili: uniform DKM, uniform Gini-indeks, uniform informacijski prispevek in uniform klasifikacijsko točnost. Mere, ki izbirajo atribute na podlagi razdalje med porazdelitvami vrednosti razredov glede na uporabljeni atribut so: Hellingerjeva razdalja, razdalja AUC, kotna razdalja in Evklidska razdalja, povzete so po [11].

Vse naštete mere so implementirane v paketu CORElearn za statistični program R. Za opisovanje mer za ocenjevanje atributov bomo v tem poglavju uporabljali naslednji zapis:

- množica atributov $A = \{A_i, i = 0...a\}$;
- za vsak diskretni atribut A_i imamo množico možnih vrednosti $\mathcal{V}_i = \{V_1, ..., V_{m_i}\}$;
- za vsak zvezni atribut A_i imamo interval možnih vrednosti $\mathcal{V}_i = [Min_i, Max_i]$, kjer je Min_i najmanjša in Max_i največja vrednost atributa A_i;

- razred je podan z atributom A_0: če rešujemo klasifikacijski problem, potem je A_0 diskretni atribut, če pa rešujemo regresijski problem, je A_0 številski atribut;
- učni primer je vektor vrednosti atributov $u_j = \langle r^{(j)}, v^{(1,j)}, ..., v^{(a,j)} \rangle$, pri tem je razred označen z $r^{(j)} = v^{(0,j)}$;
- množica učnih primerov je podana kot množica vektorjev $\mathcal{U} = \{u_j, j = 1...n\}$;
- n - število učnih primerov;
- $n_{k.}$ - število učnih primerov iz razreda r_k;
- $n_{.j}$ - število učnih primerov z j-to verjetnostjo danega atributa A;
- n_{kj} - število učnih primerov iz razreda r_k in z j-to verjetnostjo danega atributa A.

Vpeljemo še aproksimacije verjetnosti iz učne množice primerov:

- $p_{kj} = n_{kj}/n$;
- $p_{k.} = n_{k.}/n$;
- $p_{.j} = n_{.j}/n$;
- $p_{k|j} = p_{kj}/p_{.j} = n_{kj}/n_{.j}$

in naslednje entropije:
H_R - entropija razredov:

$$H_R = -\sum_k p_{k.} \log p_{k.}$$

H_A - entropija vrednosti danega atributa:

$$H_A = -\sum_j p_{.j} \log p_{.j}$$

H_{RA} - entropija produkta dogodkov razred-vrednost atributa:

$$H_{RA} = -\sum_k \sum_j p_{kj} \log p_{kj}$$

$H_{R|A}$ - pogojna entropija razreda pri dani vrednosti atributa:

$$H_{R|A} = H_{RA} - H_A = -\sum_j p_{.j} \sum_k \frac{p_{kj}}{p_{.j}} \log \frac{p_{kj}}{p_{.j}}$$

$$H_{R|A} = -\sum_j p_{.j} \sum_k p_{k|j} \log p_{k|j}$$

Ker velja $H_X \geq 0$, velja $H_{RA} \geq H_{R|A}$, prav tako veljata relaciji $H_{RA} > H_R$ in $H_{RA} > H_A$.

2.1 ReliefF

Sprva je bil algoritem RELIEF (Kira in Rendell, 1992) razvit za ocenjevanje atributov v dvorazrednih klasifikacijskih problemih. Osnovna ideja algoritma je, da za vsak učni primer poišče najbližji primer iz istega razreda in najbližji primer iz nasprotnega razreda. Na ta način oceni kvaliteto atributa glede na lokalne značilnosti razločevanja razredov. Lokalnost vključuje v oceno tudi ostale atribute in s tem RELIEF implicitno ocenjuje atribute v odvisnosti od ostalih atributov.

Izboljšana verzija algoritma RELIEF je ReliefF (Kononenko, 1994). ReliefF se od ostalih mer za ocenjevanje atributov razlikuje po tem, da ne predpostavlja apriorne in pogojne neodvisnosti atributov pri danem razredu. V splošnem ReliefF meri kvaliteto atributov glede na njihovo sposobnost razlikovanja med podobnimi primeri. Poleg tega pri ocenjevanju atributov upošteva neznane vrednosti atributov, obravnave šumne podatke in večrazredne probleme. Na šumnih podatkih RelifF poišče k najbližjih zadetkov in k najbližjih pregreškov ter upošteva povprečje njihovih prispevkov. Tipična vrednost parametra $k = 5, ..., 10$. Pri večrazrednih problemih ReliefF poišče po k najbližjih sosedov iz vsakega razreda. Prispevki posameznih razredov so uteženi z apriornimi verjetnostmi razredov.

ReliefFexpRank, ki smo ga uporabljali v naših testih je ReliefF, kjer k najbližjim primerom eksponentno pada utež z rastočim rangom. Rang primerov je določen z Manhattansko razdaljo od izbranega primera.

2.2 Informacijski prispevek

Klasična mera za pomembnost atributa je informacijski prispevek (*information gain*). Informacijski prispevek atributa je definiran kot prispevek informacije atributa k določitvi vrednosti razreda:

$$Gain(A) = H_R + H_A - H_{RA} = H_R - H_{R|A}$$

Ker je entropija mera nečistoče, veljajo za informacijski prispevek lastnosti, ki veljajo za te vrste funkcij pomembnosti atributa:

$$0 \leq Gain(A) \leq H_R$$

Če atributu A neko vrednost razbijemo na dve vrednosti in tako dobimo A', velja:

$$Gain(A') \geq Gain(A)$$

2.3 MDL

Po principu najkrajšega opisa (*Minimum Description Length*) velja, da je atribut tem pomembnejši, čimbolj je kompresiven. To pomeni, da izberemo model, ki nam da najbolj kompresiven opis podatkov in modela samega.

Imejmo problem prenosa podatkov po komunikacijskem kanalu. Oba, pošiljatelj in prejemnik, poznata vse možne vrednosti $\mathcal{V}_i$ danega atributa A, število možnih razredov m_0 in za vsak primer u_i poznata vrednost atributa $v^{(i)}$. Samo pošiljatelj pozna pravilne razrede za vse primere. Naloga pošiljatelja je poslati informacijo o razredih za vse primere, tako da je sporočilo čim krajše, t.j. kodirano s čim manjšim številom bitov.

Število bitov, ki so potrebni za zakodiranje razredov primerov z dano verjetnostno porazdelitvijo, lahko aproksimiramo z entropijo H_R, pomnoženo s številom primerov n, čemur prištejemo število bitov, potrebnih za zakodiranje porazdelitve po razredih. Ker imamo n primerov in m_0 razredov, je število različnih porazdelitev enako

$$\binom{n+m_0-1}{m_0-1},$$

kar pomeni, da je vseh možnih mej $n+m_0-1$, izbrati pa moramo m_0-1 mej. Na ta način dobimo oceno števila bitov, ki jih potrebujemo za kodiranje razredov vseh n primerov:

$$Prior\_MDL' = nH_R + \log\binom{n+m_0-1}{m_0-1}$$

Ocena števila bitov za kodiranje razredov po posameznih vrednostih atributa je vsota takih členov po vseh vrednostih atributa A:

$$Post\_MDL'(A) = nH_{R|A} + \sum_j \log\binom{n_{.j}+m_0-1}{m_0-1}$$

Pomembnost atributa $MDL'(A)$ definiramo kot kompresivnost atributa, t.j. kot razliko dolžin kode brez in z uporabo atributa, normalizirano s številom učnih primerov:

$$MDL'(A) = \frac{Prior\_MDL' - Post\_MDL'(A)}{n}$$

Če poznamo število besed (učnih primerov) lahko uporabimo optimalnejše kodiranje, namesto H_R, ki definira optimalno kodiranje samo, če je število zakodiranih besed poljubno, izračunamo število vseh možnih klasifikacij n učnih primerov:

$$\binom{n}{n_{1.},...,n_{m_0}}$$

in tako dobimo:

$$Prior\_MDL = \log\binom{n}{n_{1.},...,n_{m_0}} + \log\binom{n+m_0-1}{m_0-1}$$

$$Post\_MDL(A) = \sum_j \log \binom{n_{\cdot j}}{n_{1j}, ..., n_{m_0 j}} + \sum_j \log \binom{n_{\cdot j} + m_0 - 1}{m_0 - 1}$$

in

$$MDL(A) = \frac{Prior\_MDL - Post\_MDL(A)}{n}$$

2.4 Gini-indeks

Še ena izmed klasičnih mer za gradnjo odločitvenih modelov je Gini-indeks (*Gini-index*). Apriorni Gini-indeks je mera nečistoče, definirana z:

$$Gini\_prior = \sum_k \sum_{l \neq k} p_{k\cdot} p_{l\cdot} = 1 - \sum_k p_{k\cdot}^2$$

Pomembnost atributa $Gini(A)$ je definirana kot razlika med apriornim in pričakovanim aposteriornim Gini-indeksom:

$$Gini(A) = \sum_j p_{\cdot j} \sum_k p_{k|j}^2 - \sum_k p_{k\cdot}^2$$

$Gini(A)$ ima lastnosti, ki izvirajo iz definicije mere nečistoče:

Negativnost: $Gini(A) \geq 0$

Maksimalna vrednost:

$$Gini(A) = Gini\_prior \iff \forall j : \exists! k : p_{k|j} = 1$$

Večanje števila vrednosti atributa: $Gini(A)$ kvečjemu naraste. Ta lastnost je nezaželena, saj $Gini(A)$ precenjuje večvrednostne atribute.

2.5 Klasifikacijska točnost

Klasifikacijska točnost (*accuracy*) ima kot mera za ocenjevanje atributov precej slabosti (Brodley, 1995) v smislu strukture odločitvenega drevesa, lahko

pa je uporabna pri optimizaciji klasifikacijske točnosti odločitvenega drevesa, če rešimo problem odvisnosti atributov.

Če klasificiramo z večinskim razredom, je klasifikacijska točnost kot mera za ocenjevanje atributov definirana kot:

$$Acc = \sum_{V} P(V) \max_{c} P(C|V) \tag{2.1}$$

Če želimo izračunati klasifikacijsko točnost za binarno razdelitev večvrednostnih atributov, grupiramo vrednosti atributov v dve skupini (levo in desno) in uporabimo verjetnost skupine v računu 2.1.

2.6 DKM

Mera DKM (Dietterich, Kearns, in Mansour; 1996) deluje podobno kot informacijski prispevek. Razlika je v tem, da ima DKM drugačen kriterij za delitev atributov. Mera nečistoče distribucije razredov pri DKM je binarna entropijska funkcija definirana kot:

$$G(q) = \sqrt[2]{q(1-q)},$$

kjer q predstavlja verjetnost enega od razredov. DKM je kot mera za ocenjevanje atributov definirana kot:

$$DKM = \sum_{V} P(V)G(q)$$

2.7 Mere z enakomerno porazdelitvijo

Uporabljene mere z enakomerno apriorno porazdelitvijo ($uniform$) so: uniform DKM, uniform Gini-indeks, uniform informacijski prispevek in uniform klasifikacijska točnost. Delujejo podobno kot njihovi soimenjaki (DKM, Gini-indeks, informacijski prispevek in klasifikacijska točnost) brez predpone, z razliko, da vsilijo enakomerno porazdelitev pred razbitjem primerov glede na dani atribut.

Naj bo c_i i-ti razred in v_j j-ta veja drevesa. Enakomerno porazdelitev pred razbitjem primerov dosežemo z izbiro takšnega faktorja $\alpha > 0$ za vsak razred c_i, da je

$$n'_{c_i,v_j} = \alpha_i n_{c_i,v_j}$$

in je

$$\sum n'_{c_i} = \alpha_i n_{c_i}$$

enaka za vse razrede c_i.

2.8 Hellingerjeva razdalja

Razdaljo med dvema diskretnima distribucijama $P = \{p_i\}_{i=1}^{s}$ in $Q = \{q_i\}_{i=1}^{s}$ v vejah drevesa nad enako diskretno domeno z s elementi z verjetnostmi p_i in q_i za $i = 1, ..., s$ definiramo kot:

$$Hellinger(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{s}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

Faktor $\frac{1}{2}$ je vključen zaradi normalizacije, da je maksimalna razdalja med dvema distribucijama največ 1.

2.9 Razdalja AUC

Predpostavimo, da je diskretna statistika A z s vrednostmi $a_1, ..., a_s$ in distribucijo $P = \mathrm{P}(A|c_1)$ za razred c_1 in $Q = \mathrm{P}(A|c_2)$ za razred c_2 edini podatek, ki ga imamo na voljo za razločevanje razredov c_1 in c_2. Potem je razločevanje mogoče le, če sta P in Q različna. Razdaljo med distribucijama P in Q izmerimo s pomočjo AUC (*Area Under the ROC Curve*) množice optimalnih klasifikatorjev za različne matrike napak. Za optimalno izbiro množice klasifikatorjev je $AUC \geq 1/2$.

Da bosta razdalji dveh identičnih distribuciji P in Q enaki 0, je AUC razdalja definirana kot:

$$dist_{AUC}(P,Q) = 2AUC(P,Q) - 1$$

to razdaljo pa ocenimo z:

$$dist_{AUC}(P,Q) = \sum_{1 \le i \le j \le s} |p_i q_j - p_j q_i|$$

2.10 Kotna razdalja

Naj bo φ kot med dvema vektorjema $P = \{p_i\}_{i=1}^s$ in $Q = \{q_i\}_{i=1}^s$. Potem je kotna razdalja (*Distance Angle*) dana kot:

$$CosAngleDistance(P,Q) = \sqrt{1 - \cos\varphi} = \sqrt{\frac{1}{2}\sum_{i=1}^{s}\left(\frac{p_i}{\sqrt{\sum_j p_j^2}} - \frac{q_i}{\sqrt{\sum_j q_j^2}}\right)^2}$$

Faktor $\frac{1}{2}$ je vključen zaradi normalizacije, da je maksimalna razdalja med dvema kotoma največ 1.

2.11 Evklidska razdalja

Mera *DistEuclid*, kot že ime pove, za ocenjevanje atributov uporablja Evklidsko razdaljo. Slednja je za dve distribuciji $P = \{p_i\}_{i=1}^s$ in $Q = \{q_i\}_{i=1}^s$ definirana kot:

$$Euclidean(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{s}(p_i - q_i)^2}$$

pri čemer faktor $\frac{1}{2}$ zagotovi, da je maksimalna razdalja med dvema distribucijama največ 1.

Poglavje 3

Metodologija testiranja

Za testiranje opisanih mer za ocenjevanje atributov smo potrebovali ustrezne neuravnotežene dvorazredne množice podatkov. Iz UCI ML Repository [5] smo izbrali 15 domen, iz LIBSVM [1] 4 domene in iz Delve [7] repozitorija 1 domeno. Skupaj smo tako imeli 20 domen, ki jih prikazuje tabela 3.1. Domene smo uredili tako, da smo vsem razrede binarizirali in tam, kjer je bilo razredov več, logično združili razrede tako, da smo dobili povsod dvorazredne množice.

Pripravljene domene smo vzorčili in z njimi zgradili odločitvena drevesa z opisanimi merami. Modele smo testirali s sorazmernim prečnim preverjanjem reda 5×2 in izmerili AUC zgrajenih modelov za vsako naravno in vzorčeno domeno. Uspešnost mer smo med sabo primerjali s Friedmanovim testom in tam, kjer smo zavrgli ničelno hipotezo o enakosti, nadaljevali z Nemenyievim testom. Domene smo pripravili s programom Microsoft Excel; vzorčenje, gradnjo odločitvenih modelov in Friedmanov test smo realizirali v programu R. S programom Orange in programskim jezikom Python smo izvedli Nemenyiev test in grafično predstavili, kje se pari uporabljenih mer razlikujejo.

Z	Domena	Izvor	I	A	D	C	MC (%)
1	adult	UCI ML	48842	14	8	6	24
2	australian-credit	UCI ML	690	14	8	6	44
3	blood-transfusion	UCI ML	748	4	0	4	24
4	cod-rna	LIBSVM	59535	8	0	8	33
5	contraceptive	UCI ML	1473	9	7	2	43
6	fourclass	LIBSVM	862	2	0	2	36
7	gamma-telescope	UCI ML	19020	10	0	10	35
8	german-credit	UCI ML	1000	20	13	7	30
9	letter-recognition	UCI ML	20000	16	0	16	19
10	mammographic-mass	UCI ML	961	5	4	1	46
11	musk-v2	UCI ML	6598	166	0	166	15
12	page-blocks	UCI ML	5473	10	0	10	10
13	pima-diabetes	UCI ML	768	8	0	8	35
14	segmentation	UCI ML	2310	18	0	18	43
15	spam	UCI ML	4601	57	0	57	39
16	splice	UCI ML	1535	60	60	0	50
17	svmguide1	LIBSVM	3089	4	0	4	35
18	svmguide3	LIBSVM	1243	21	0	21	24
19	tic-tac-toe	UCI ML	958	9	9	0	35
20	titanic	Delve	2201	3	3	0	32

Tabela 3.1: Izbrane domene; Z označuje zaporedno številko domene, I število primerov v domeni, A število atributov v domeni (brez razreda), D število diskretnih atributov, C število številskih atributov in MC (%) označuje delež manjšinskega razreda v domeni.

3.1 Priprava domen

Domeni contraceptive smo razrede: 1 — ni bila uporabljena kontracepcija, 2 — kratkoročno jemanje kontracepcije in 3 — dolgoročno jemanje kontracepcije, združili in diskretizirali v dva razreda: 1 v n ter 2 in 3 v y, torej n — ni bila uporabljena kontracepcija in y — je bila uporabljena kontracepcija. Domeni letter-recognition smo razrede, ki so bile črke angleške abecede, združili v razreda *sa* kamor smo združili samoglasnike ter *so* kamor smo združili preostale črke. Prav tako smo domeni page-blocks razrede $h.line$, $v.line$, $graphic$ in $picture$ združili v razred g — grafika, razred $text$ pa je postal t.

Domeni segmentation smo razrede sky, $foliage$, $path$ in $grass$ združili v razred $nature$, razrede $brickface$, $window$ in $cement$ pa v $building$. Poleg tega smo odstranili atribut $region-pixel-count$, ker so bil vse vrednosti enake. Pri domeni splice smo zavrgli razred $Neither$ in tako sta ostala le razreda EI in IE. Domeni musk-v2 smo odstranili atribut $conformation\_name$ — simbolična imena enaka kot razredi, ker ni imeli nobenega klasifikacijskega pomena.

Manjkajoče vrednosti v domenah adult, mammographic-mass, pima-diabetes in spam smo nadomestili z NA. Vse ostale domene so bile že dvorazredne množice. Pretvorbe razredov prikazuje tabela 3.2.

3.2 Vzorčenje

Da smo dobili različne stopnje neuravnoteženosti množic smo izvedli vzorčenje primerov s pomočjo programa R. Domene smo vzorčili, da smo dobili razmerja razredov 1:5, 1:10, 1:50, 1:100, 1:500 in 1:1000.

Vzorčili smo tako, da smo sprva iz množice manjšinskega razreda naključno izbrali 50 primerov, nato pa iz množice večinskega razreda zavrgli ali razmnožili r primerov, da smo dosegli željeno razmerje. Psevdokodo algoritma za vzorčenje prikazuje algoritem 1.

Domena	Razred
adult	>50K, <=50K
australian-credit	1=y, 0=n
blood-transfusion	1=y, 0=n
cod-rna	+1=p, −1=n
contraceptive	2, 3=y, 1=n
fourclass	+1=p, −1=n
gamma-telescope	g, h
german-credit	1=y, 2=n
letter-recognition	a, e, i, o, u=sa, ostali=so
mammographic-mass	benign, malignant
musk-v2	musk=m, non-musk=nm
page-blocks	text=t,
	h.line, v.line, graphic, picture=g
pima-diabetes	1=y, 0=n
segmentation	sky, foliage, path, grass=nature,
	brickface, window, cement=building
spam	1=y, 0=n
splice	EI, IE (razred N zavržen)
svmguide1	1=p, 0=n
svmguide3	1=p, 0=n
tic-tac-toe	positive, negative
titanic	yes, no

Tabela 3.2: Pretvorjeni razredi.

Algoritem 1. Vzorčenje podatkov do razmerja.

```
1: procedure doSample(data, razmerje)
2:     minSubset ← manjšinski razred v data
3:     maxValue ← # primerov večinskega razreda v data
4:     maxSubset ← večinski razred v data
5:     minC ← naključno izberi 50 primerov iz minSubset
6:     r ← (50 × razmerje) − maxValue
7:     if r < 0 then
8:         zavrzi r primerov iz maxSubset
9:     else
10:        razmnoži maxSubset za r primerov
11:    end if
12:    dataS ← minC + maxSubset
13:    return dataS
14: end procedure
```

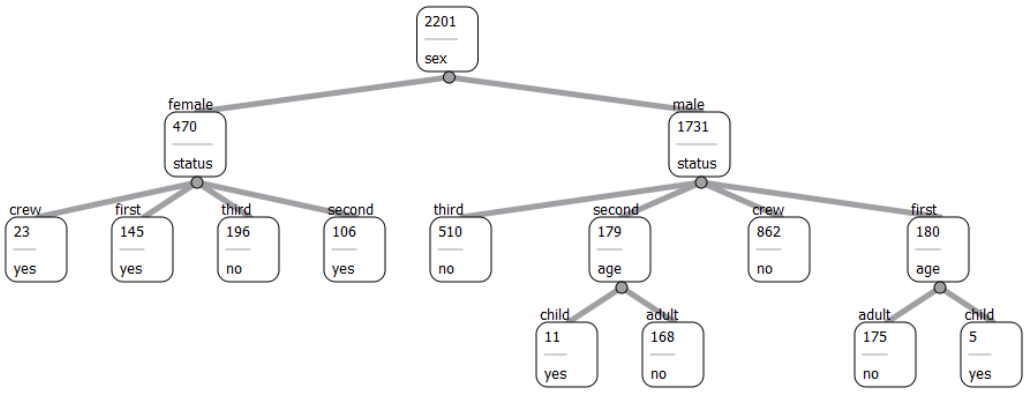
3.3 Odločitvena drevesa

Kot modele smo uporabili odločitvena drevesa, zgrajena z vsako od opisanih 14 mer za ocenjevanje atributov.

Odločitveno drevo (*decision tree*) je sestavljeno iz notranjih vozlišč, ki ustrezajo atributom, vej, ki ustrezajo podmnožicam vrednosti atributov, in listov, ki ustrezajo razredom. Ena pot v drevesu od korena do lista ustreza enemu odločitvenemu pravilu. Primer odločitvenega drevesa prikazuje slika 3.1.

3.4 Sorazmerno prečno preverjanje reda 5×2

Zgrajene odločitvene modele smo testirali s sorazmernim prečnim preverjanjem reda 5×2 (*stratified* 5×2 *cross-validation*). Pri sorazmernemu prečnemu preverjanju ohranjamo približno enako distribucijo razredov v vseh podmnožicah. Na ta način je distribucija razredov v vsaki učni in testni množici



Slika 3.1: Primer odločitvenega drevesa zgrajenega z mero Gini-indeks na domeni titanic. Zgornje število v vozlišču pomeni število primerov, spodaj je atribut glede na katerega delimo. Listi drevesa predstavljajo razrede. Namen drevesa je klasificirati, ali je oseba na Titaniku preživela.

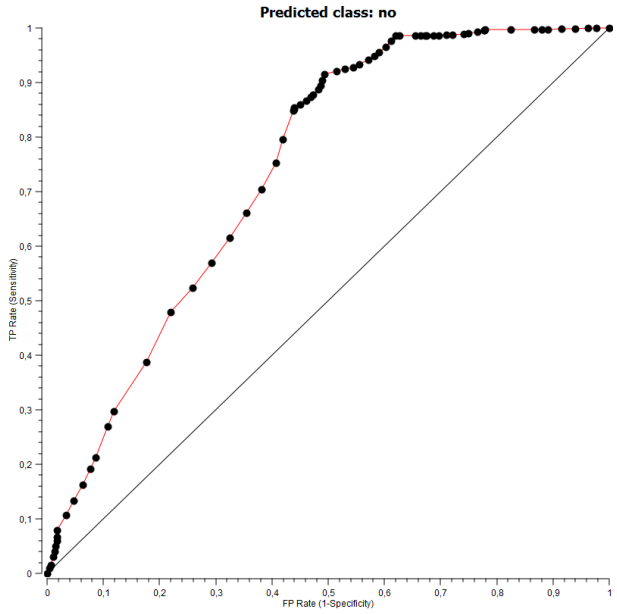
približno enaka. Psevdokodo algoritma sorazmernega prečnega preverjanja reda 5×2 prikazuje algoritem 2.

3.5 AUC

Uspešnost uporabljenih mer smo ocenili z AUC (*Area Under the ROC Curve*), to je ploščino pod ROC (*Reciver Operating Characteristic*) krivuljo. ROC krivulja (slika 3.2) omogoča analizo razmerja med senzitivnostjo (verjetnost, da klasifikator zazna pozitivni primer — razred c_p v dvorazredni množici) in specifičnostjo (verjetnost, da klasifikator zazna negativni primer — razred c_n v dvorazredni množici). Neka mera je boljša od druge, če ima večjo vrednost AUC. AUC je enaka verjetnosti, da bo mera pravilno razločila med pozitivnimi in negativnimi primeri. Klasifikator, ki vse primere pravilno klasificira, ima oceno AUC enako 1, tisti ki jih klasificira naključno pa 0.5.

Algoritem 2. Sorazmerno prečno preverjanje reda 5×2.

1: $data \leftarrow$ preberi domeno
2: $minValue \leftarrow$ # primerov manjšinskega razreda v $data$
3: $minSubset \leftarrow$ manjšinski razred v $data$
4: $maxValue \leftarrow$ # primerov večinskega razreda v $data$
5: $maxSubset \leftarrow$ večinski razred v $data$
6: **for** $i = 1$ **to** 5 **do**
7: $min \leftarrow$ naključno izberi $minValue/2$ primerov iz $minSubset$
8: $max \leftarrow$ naključno izberi $maxValue/2$ primerov iz $maxSubset$
9: $ucna \leftarrow min + max$
10: $testna \leftarrow \neg ucna$
11: $ucnaS \leftarrow doSample(ucna, razmerje)$
12: $testnaS \leftarrow doSample(testna, razmerje)$
13: **for** $j = 1$ **to** 2 **do**
14: zgradi odločitveni model z $ucnoS$ in ga testiraj s $testnoS$ množico
15: zamenjaj $ucnoS$ in $testnoS$ množico
16: **end for**
17: **end for**



Slika 3.2: Primer ROC krivulje za klasifikacijsko drevo 3.1 za razred "niso preživeli". Os x prikazuje relativno število napačno klasificiranih negativnih primerov (1-specifičnost), os y prikazuje relativno število pravilno klasificiranih pozitivnih primerov (senzitivnost). Vsak klasifikator je na sliki prikazan kot točka na ROC krivulji. Točka nad diagonalo predstavlja sprejemljivo klasifikacijo, točka pod diagonalo pa slabo klasifikacijo.

3.6 Statistični testi

Ker smo testirali več mer za ocenjevanje atributov na večih domenah, smo za analizo, ali se mere med seboj razlikujejo, uporabili Friedmanov test [3]. Glede na rezultat Friedmanovega testa smo, da bi ugotovili, kateri pari mer se medseboj statistično značilno razlikujejo, uporabili Nemenyiev test. Oba testa opisujemo v nadaljevanju.

Friedmanov test

Friedmanov test je neparametrični test, ki rangira uspešnost vsakega algoritma na vsaki domeni posebej. Najboljši algoritem dobi rang 1, drugi najboljši 2, itd. V primerih, ko je več algoritmov enako uspešnih, se jim dodeli povprečni rang. Če imamo k algoritmov in D domen in je r_i^j rang j-tega algoritma na i-ti domeni, je povprečni rang definiran kot:

$$R_j = \frac{1}{D} \sum_{i=1}^{D} r_i^j$$

Pri Friedmanovem testu najprej postavimo ničelno hipotezo H_0 in alternativno hipotezo H_a ter določimo stopnjo zaupanja α.

$$H_0 = \textit{algoritmi se ne razlikujejo med sabo}$$
$$H_a = \textit{algoritmi se razlikujejo med sabo}$$
$$\alpha = 0.05$$

Izračunamo Friedmanovo statistiko:

$$\chi_F^2 = \frac{12D}{k(k+1)} \left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right)$$

Ker pri velikemu številu primerjav obstaja nevarnost, da so razlike zgolj naključje, pravimo, da je Friedmanova statistika konzervativna. Zato izračunamo izboljšano Friedmanovo statistiko porazdeljeno po zakonu F s $(k-1)$ in $(k-1) \times (D-1)$ prostostnimi stopnjami:

$$F_F = \frac{(D-1)\chi_F^2}{D(k-1) - \chi_F^2}$$

Hipotezo H_0 zavržemo, če je $F_F > F_{(k-1),(k-1)\times(D-1);\alpha}$.

Nemenyiev test

Nemenyiev test uporabimo, kadar zavržemo ničelno hipotezo H_0, in želimo primerjati vsak algoritem z vsakim. Z Nemenyievim testom želimo ugotoviti, kateri pari algoritmov se med seboj statistično značilno razlikujejo.

V ta namen izračunamo kritično razdaljo CD, ki pove, za koliko se morata dva algoritma razlikovati, da je razlika značilna:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6D}}$$

Pri $k = 14$ in $\alpha = 0.05$ je kritična vrednost $q_\alpha = 3.354$. Dva algoritma j_1 in j_2 se značilno razlikujeta, če je razlika med povprečnima rangoma večja ali enaka kritični razdalji CD:

$$|R_{j_1} - R_{j_2}| \geq CD$$

3.7 Uporabljeni programi

Pri delu smo uporabljali program Microsoft Excel za pripravo domen, statistični program R s paketom CORElearn za vzorčenje množic, gradnjo odločitvenih modelov in za računanje Friedmanove statistike. Program Orange ter programskim jezik Python smo uporabili za izvedbo in grafično predstavitev Nemenyievega testa.

R

R [8] je programski jezik in odprtokodno statistično okolje, namenjeno statistični obdelavi podatkov in grafičnim izrisom. Vključuje veliko zbirko orodij za obdelavo podatkov, metode za izvajanje računskih operacij nad tabelami in matrikami ter grafična orodja za analizo in prikaz podatkov. Ker je R tudi programski jezik, omogoča pisanje skript — vključuje pogojne operatorje, zanke in dostope do vhodno / izhodnih naprav. R je enostavno razširljiv s paketi, ki so na voljo preko CRAN repozitorija.

CORElearn [10] je razširitveni paket za R, ki omogoča rudarjenje podatkov. Vsebuje različne klasifikacijske in regresijske modele ter algoritme za izbiro in vrednotenje atributov ter modelov.

Orange

Orange [2] je brezplačen program za strojno učenje in podatkovno rudarjenje. Razvit je bil v Laboratoriju za bioinformatiko na Fakulteti za računalništvo in informatiko v Ljubljani. Orange vključuje vrsto tehnik za upravljanje in procesiranje podatkov, nadzorovano in nenadzorovano učenje, analizo zmogljivosti ter vrsto tehnik za vizualizacijo in predstavitev tako podatkov kot modelov. Uporablja se lahko preko Orange Canvas grafičnega uporabniškega vmesnika ali pa preko skript v jeziku Python.

Poglavje 4

Rezultati in ugotovitve

V tem poglavju prikazujemo rezultate testiranj. Rezultati so prikazani glede na razmerja razredov v domenah. Na koncu poglavja so podane sklepne ugotovitve.

Ker je iz samih ocen AUC le težko primerjati mere, so v tabeli 4.2 prikazani povprečni rangi mer glede na razmerja razredov v domenah, tabele ocen AUC pa so v prilogi B. Značilne razlike med merami po Nemenyievem testu so prikazane z grafi kritične razdalje. Skupine mer, ki se med sabo značilno ne razlikujejo, so povezane.

Ničelno hipotezo smo zavrnili pri vseh razmerjih, torej se je povsod delovanje mer statistično značilno razlikovalo. Zavrnitveni kriterij Friedmanovega testa in kritično razdaljo Nemenyievega testa prikazuje tabela 4.1.

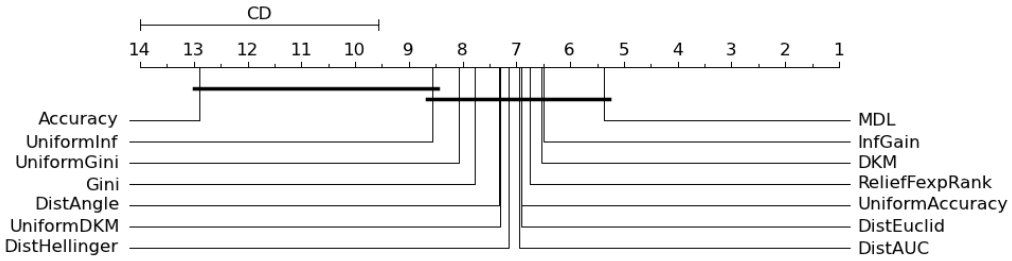
Razmerje	K	D	F	CD
do 1:100	14	20	1.76	4.437
1:500	14	19	1.76	4.552
1:1000	14	18	1.76	4.677

Tabela 4.1: Zavrnitveni kriterij Friedmanovega in kritična razdalja Nemenyievega testa. K predstavlja število uporabljenih mer in D število uporabljenih domen. F pomeni F-distribucija s $(K-1)$ in $(K-1)\times(D-1)$ prostostnimi stopnjami. CD je kritična razdalja po Nemenyievem testu.

Razmerje	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
Naravno	6.75/4.83	6.5/3.49	5.38/4.06	7.78/3.93	12.9/2.47	6.53/3.34	7.3/3.18
1:5	6.9/4.67	6.75/4	6.62/3.97	8.72/3.74	12.55/3.23	7.83/3.41	8.35/3.55
1:10	5.22/4.53	7.05/3.39	7.33/3.84	9.38/4.18	12.9/2.61	7.5/3.55	8.35/3.35
1:50	6.35/4.55	7.15/3.4	7.2/3.75	8.4/4.54	12.4/3.38	6.38/3.66	7.2/3.53
1:100	5.25/4.72	4.95/3.94	5.95/4.78	8.75/4.71	13/2.28	6.2/3.01	6.92/3.01
1:500	6.68/3.15	4.47/3.49	7.03/4.84	9.21/4.79	12/3.71	6/3.35	6.84/3.22
1:1000	7.86/4.62	5.11/4.34	6.25/4.7	8.5/4.9	11.69/3.59	5.83/3.39	6.17/3.08

Razmerje	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
Naravno	8.07/3.13	8.57/3.23	6.9/3.95	7.15/3.1	6.95/3.89	7.33/4.03	6.9/3.72
1:5	8.1/3.4	8.2/3.38	6/2.98	7.78/3.5	4.6/2.58	6.9/4.4	5.7/3.45
1:10	7.78/3.35	6.65/4.07	5.85/3.1	8.2/3.21	6.1/3.05	6.78/3.73	5.92/3.55
1:50	7.83/3.67	8.22/3.77	6.53/3.71	6.42/3.57	7.05/3.02	7.45/4.02	6.42/3.98
1:100	7.12/3.82	8.28/3.54	8.43/2.85	6.22/2.67	7.97/2.41	8.47/3.33	7.47/3.3
1:500	7.5/4	8.08/3.59	6.89/3.34	6.13/3.01	7.74/3.66	9.34/3.1	7.08/3.58
1:1000	7.44/2.54	7.19/3.76	8.94/2.58	6.22/3.23	7.47/2.8	8.72/3.5	7.58/3.63

Tabela 4.2: Povprečni rangi in standardni odklon povprečnih rangov mer, glede na razmerje razredov v domenah.



Slika 4.1: Graf kritične razdalje glede na AUC rangiranih mer pri naravnem razmerju razredov.

4.1 Naravno razmerje

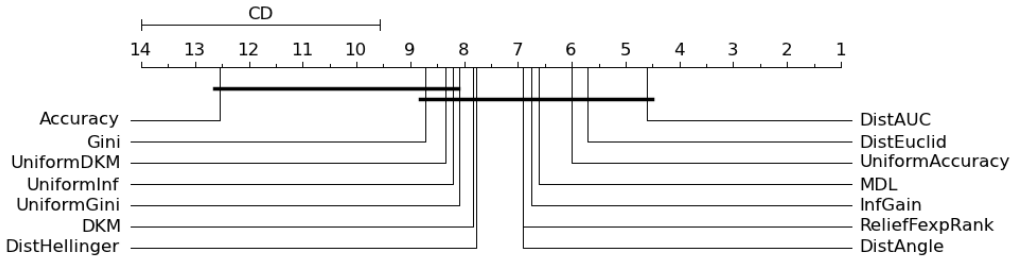
Slika 4.1 prikazuje značilne razlike mer pri naravnem razmerju porazdelitve razredov v domenah. Značilno se razlikuje delovanje klasifikacijske točnosti in ostalih mer razen uniform informacijskega prispevka. Čeprav se značilno ne razlikujeta klasifikacijska točnost in uniform informacijski prispevek je velika razlika med njunima povprečnima rangoma. Klasifikacijska točnost dosega povprečni rang 12.9 in odklon 2.47, uniform informacijski prispevek pa 8.57 in odklon 3.23.

Kot najboljša mera se izkaže MDL s povprečnim rangom 5.38 in odklonom 4.06, kot najslabša pa klasifikacijska točnost. Klasifikacijska točnost kaže glede na povprečni rang najmanjši odklon, kar pomeni, da je pri večini domen dosegala najslabši AUC. Največji odklon kaže mera ReliefF s povprečnim rangom 6.75 in odklonom 4.83.

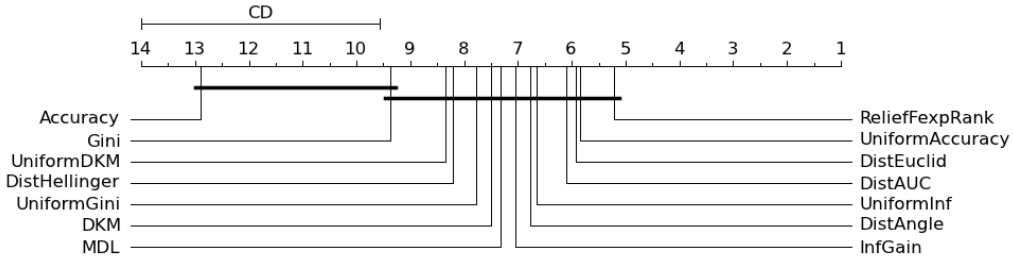
4.2 Razmerje 1:5

Značilne razlike mer pri razmerju razredov 1:5 prikazuje Slika 4.2. Značilno se ne razlikuje delovanje mer klasifikacijska točnost, Gini-indeks, uniform DKM, uniform informacijski prispevek in uniform Gini-indeks. Značilne razlike obstajajo med ostalimi merami in klasifikacijsko točnostjo.

Najbolje se odreže mera razdalja AUC s povprečnim rangom 4.6 in od-



Slika 4.2: Graf kritične razdalje glede na AUC rangiranih mer pri razmerju razredov 1:5.



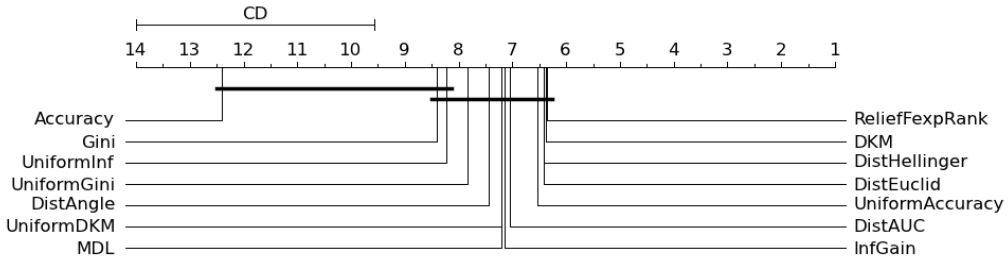
Slika 4.3: Graf kritične razdalje glede na AUC rangiranih mer pri razmerju razredov 1:10.

klonom 2.58, najslabše pa ponovno klasifikacijska točnost z rangom 12.55 in odklonom 3.23. Najmanjši odklon kaže razdalja AUC, največjega pa ponovno ReliefF z rangom 6.9 in odklonom 4.67.

4.3 Razmerje 1:10

Slika 4.3 prikazuje značilne razlike mer pri razmerju porazdelitve razredov 1:10. Značilno se razlikuje delovanje mere klasifikacijska točnost in ostalih mer razen mere Gini-indeks.

Najboljša mera je ReliefF s povprečnim rangom 5.22 in odklonom 4.53, najslabša pa klasifikacijska točnost s povprečnim rangom 12.9 in odklonom 2.61. Najmanjši odklon kaže klasifikacijska točnost največjega pa ReliefF.



Slika 4.4: Graf kritične razdalje glede na AUC rangiranih mer pri razmerju razredov 1:50.

4.4 Razmerje 1:50

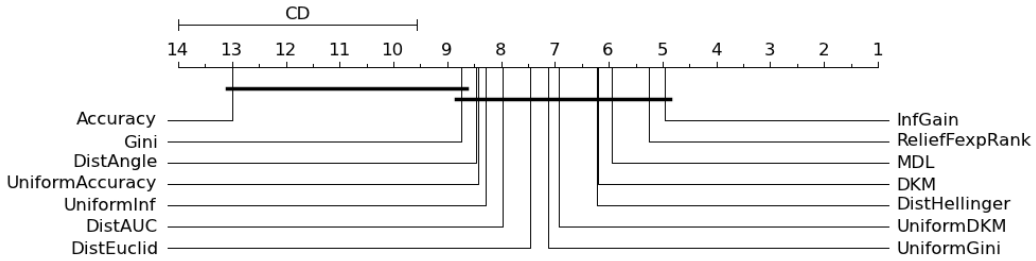
Značilne razlike mer pri razmerju razredov 1:50 prikazuje Slika 4.4. Značilna razlika obstaja med merami klasifikacijska točnost in ostalimi brez mer Gini-indeks in uniform informacijski prispevek.

Kot najboljša mera se izkaže ponovno ReliefF s povprečnim rangom 6.35 in odklonom 4.55. Najslabše se ponovno odreže klasifikacijska točnost s povprečnim rangom 12.4 in odklonom 3.38. Kot pri razmerju 1:10 kaže najmanjši odklon klasifikacijska točnost največjega pa ReliefF.

4.5 Razmerje 1:100

Slika 4.5 prikazuje značilne razlike mer pri razmerju porazdelitve razredov 1:100. Kot pri razmerju 1:10 tu obstajajo značilne razlike med mero klasifikacijska točnost in ostalimi brez mere Gini-indeks.

Najboljša mera je informacijski prispevek s povprečnim rangom 4.95 in odklonom 3.94, najslabša pa klasifikacijska točnost s povprečnim rangom 13 in odklonom 2.28. Največji odklon kaže mera MDL z rangom 5.95 in odklonom 4.78, najmanjšega pa klasifikacijska točnost.



Slika 4.5: Graf kritične razdalje glede na AUC rangiranih mer pri razmerju razredov 1:100.

4.6 Razmerje 1:500

Pri razmerju razredov 1:500 nam ni uspelo zgraditi odločitvenih dreves za domeno spam. Posledično je kritična razdalja nekoliko večja, zavrnitveni kriterij Friedmanovega testa pa se ne razlikuje.

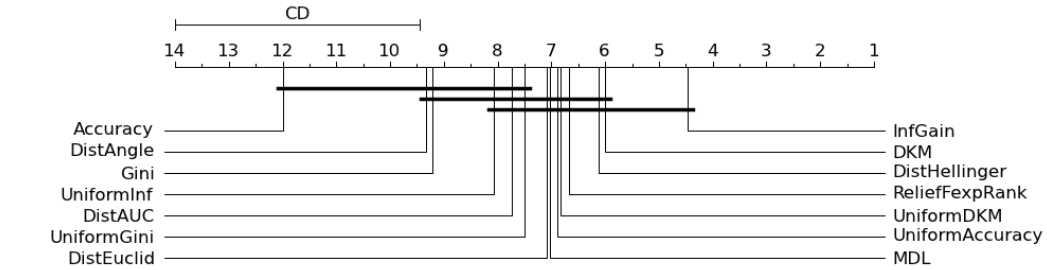
Značilne razlike mer pri razmerju razredov 1:500 prikazuje Slika 4.6. Delovanje mer se tu najbolj razlikuje in kot daleč najboljša mera se izkaže informacijski prispevek s povprečnim rangom 4.47 in odklonom 3.49. Daleč najslabša mera je ponovno klasifikacijska točnost s povprečnim rangom 12 in odklonom 3.71.

Največji odklon kaže MDL s povprečnim rangom 7.03 in odklonom 4.84, najmanjšega pa Hellingerjeva razdalja s povprečnim rangom 6.13 in odklonom 3.01.

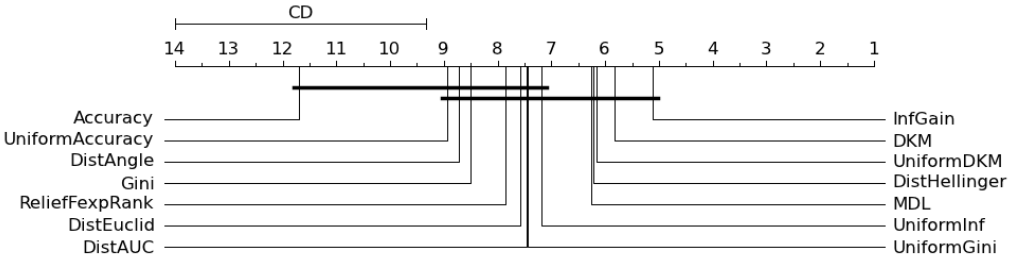
4.7 Razmerje 1:1000

Pri razmerju razredov 1:1000 nam ni uspelo zgraditi odločitvenih dreves za domeni musk-v2 in spam. Tudi tu je kritična razdalja nekoliko večja, zavrnitveni kriterij Friedmanovega testa pa se ne razlikuje.

Slika 4.7 prikazuje značilne razlike mer pri razmerju porazdelitve razredov 1:1000. Značilno se razlikuje delovanje mere klasifikacijska točnost in mer MDL, Hellingerjeva razdalja, uniform DKM, DKM in informacijskega



Slika 4.6: Graf kritične razdalje glede na AUC rangiranih mer pri razmerjem razredov 1:500.



Slika 4.7: Graf kritične razdalje glede na AUC rangiranih mer pri razmerjem razredov 1:1000.

prispevka.

Ponovno se kot najboljša mera izkaže informacijski prispevek s povprečnim rangom 5.11 in odklonom 4.34, najslabše pa klasifikacijska točnost s povprečnim rangom 11.69 in odklonom 3.59. Največji odklon kaže mera Gini-indeks s povprečnim rangom 8.5 in odklonom 4.9, najmanjšega pa uniform Gini-indeks s povprečnim rangom 7.44 in odklonom 2.54.

4.8 Primerjava klasičnih in uniform mer

V prilogi A so dani grafi, kateri prikazujejo razlike v ocenah AUC na posameznih domenah in razmerjih razredov v domenah za vsako klasično in njeno inačico z vsiljeno enakomerno porazdelitvijo (uniform).

Kakor je razvidno iz grafov, so razlike med ocenami AUC mere DKM in uniform DKM zelo majhne, skorajda neopazne. Edina opazna razlika je pri domeni spam pri razmerju razredov do 1:100 in pri domeni pima-diabetes pri razmerju razredov 1:500. Iz teh ugotovitev lahko sklepamo, da, glede na to, da uniform mere vsilijo enakomerno porazdelitev razredov, je že sama mera DKM neobčutljiva na neuravnotežene porazdelitve razredov. Zelo podobno je tudi delovanje mere DKM in Hellingerjeve razdalje, kot prikazuje slika A.8. Opazna je le razlika AUC pri testiranju na domeni spam, tako da je tudi Hellingerjeva razdalja neobčutljiva na neuravnoteženosti razredov.

Opazna izboljšava je pri primerjavi klasifikacijske točnosti in uniform klasifikacijske točnosti. Mera klasifikacijska točnost se z vsiljevanjem enakomerne porazdelitve razredov izboljša in postane ponekod primerljiva z drugimi merami, kot je razvidno iz tabele 4.2.

Delovanje mer informacijski prispevek in uniform informacijski prispevek ter Gini-indeks in uniform Gini-indeks se pri naravni porazdelitvi razredov skorajda ne razlikuje. Razlika je opazna pri večanju neuravnoteženosti porazdelitve razredov in sicer se delovanje informacijske prispevka z vsiljevanjem enakomerne porazdelitve razredov poslabša, delovanje Gini-indeksa pa izboljša, kot je razvidno iz tabele 4.2.

4.9 Ugotovitve

Pri različnih razmerjih porazdelitve razredov v domenah mere delujejo različno dobro. Pri naravnem razmerju razredov se kot najboljša mera izkaže MDL, pri lažji neuravnoteženosti razredov (1:5) pa kotna razdalja. Pri srednjih neuravnoteženostih razredov (1:10 in 1:50) se najbolje odreže ReliefF in pri večjih neuravnoteženostih razredov (1:100, 1:500 in 1:1000) se najboljše odreže informacijski prispevek. Ne glede na razmerja se mera klasifikacijska točnost izkaže povsod kot najslabša. Tabela 4.3 prikazuje povzetek delovanja mer pri različnih razmerjih porazdelitev razredov. Če primerjamo delovanje mer na vseh domenah in vseh razmerjih, se kot najboljša mera izkaže in-

Razmerje	Najboljša	Najslabša	Največji odklon	Najmanjši odklon
Naravno	MDL	Accuracy	ReliefFexpRank	Accuracy
	5.38/4.06	12.9/2.47	6.75/4.83	12.9/2.47
1:5	DistAUC	Accuracy	ReliefFexpRank	DistAUC
	4.6/2.58	12.55/3.23	6.9/4.67	4.6/2.58
1:10	ReliefFexpRank	Accuracy	ReliefFexpRank	Accuracy
	5.22/4.53	12.9/2.61	5.22/4.53	12.9/2.61
1:50	ReliefFexpRank	Accuracy	ReliefFexpRank	Accuracy
	6.35/4.55	12.4/3.38	6.35/4.55	12.4/3.38
1:100	InfGain	Accuracy	MDL	Accuracy
	4.95/3.94	13/2.28	5.95/4.78	13/2.28
1:500	InfGain	Accuracy	MDL	DistHellinger
	4.47/3.49	12/3.71	7.03/4.84	6.13/3.01
1:1000	InfGain	Accuracy	Gini	UniformGini
	5.11/4.34	11.69/3.59	8.5/4.9	7.44/2.54

Tabela 4.3: Najboljši in najslabši povprečni rang ter največji in najmanjši standardni odklon mer glede na razmerja razredov v domenah.

formacijski prispevek, tej pa sledita ReliefF in MDL. Globalni pogled na delovanje mer pri vseh razmerjih prikazuje tabela 4.4, iz katere je razvidno, da med dobrimi merami obstajajo majhne razlike.

Ugotovili smo, da mere DKM, uniform DKM in Hellingerjeva razdalja delujejo precej podobno in so neobčutljive na neuravnoteženost razredov. Mera uniform klasifikacijska točnost predstavlja veliko izboljšavo napram klasifikacijski točnosti, uniform Gini-indeks je prav tako nekoliko izboljšani Gini-indeks. Vsiljevanje enakomerne porazdelitve poslabša delovanje mere informacijski prispevek pri večji neuravnoteženosti razredov.

Mera	Rang	Razlika
InfGain	6/1.12	−
ReliefFexpRank	6.43/0.94	+0.43
MDL	6.54/0.72	+0.11
DKM	6.61/0.76	+0.07
DistEuclid	6.72/0.73	+0.11
DistAUC	6.84/1.16	+0.12
DistHellinger	6.87/0.84	+0.03
UniformAccuracy	7.08/1.18	+0.21
UniformDKM	7.3/0.8	+0.22
UniformGini	7.69/0.36	+0.39
DistAngle	7.86/0.99	+0.17
UniformInf	7.88/0.69	+0.02
Gini	8.68/0.53	+0.8
Accuracy	12.49/0.5	+3.81

Tabela 4.4: Povprečni rangi in standardni odkloni mer na vseh razmerjih.

Poglavje 5

Sklepne ugotovitve

V diplomskem delu smo analizirali delovanje mer za ocenjevanje atributov pri različnih stopnjah neuravnoteženosti porazdelitve razredov. V ta namen smo izbrali 20 realnih domen, jih priredili za testiranje in na njih izvedli vzorčenje za razmerja 1:5, 1:10, 1:50, 1:100, 1:500 in 1:1000. Na vzorčenih domenah smo z izbranimi merami za ocenjevanje atributov zgradili odločitvena drevesa in jih preverili s sorazmernim prečnim preverjanjem reda 5×2.

Dobljene ocene AUC smo med sabo primerjali z Friedmanovim testom in tam, kjer smo zavrgli ničelno hipotezo, da so vse mere enako dobre, nadaljevali z Nemenyievem testom. Statistično značilne razlike med merami po Nemenyievem testu smo grafično prikazali.

5.1 Glavne ugotovitve

Pri naravnem razmerju porazdelitve razredov se je kot najboljša mera izkazala MDL, pri razmerju 1:5 je prevladala kotna razdalja. Pri srednji neuravnoteženosti razredov, pri razmerjih 1:10 in 1:50 je bila najboljša mera ReliefF, pri večjih neuravnoteženostih pa informacijski prispevek. Klasifikacijska točnost je bila povsod najslabša s precejšnjo razliko rangov do predzadnje mere.

Delovanje mer DKM, uniform DKM in Hellingerjeva razdalja se ne raz-

likuje in vse so neobčutljive na neuravnoteženost razredov. Vsiljevanje enakomerne porazdelitve razredov izboljša delovanje klasifikacijske točnosti in nekoliko izboljša delovanje mere Gini-indeks. Informacijski prispevek se z vsiljevanjem enakomerne porazdelitve poslabša, zlasti pri večjih neuravnoteženostih razredov.

5.2 Ideje za izboljšave in nadaljnje delo

Pri velikih množicah se izkaže testiranje in gradnja odločitvenih dreves za precej časovno zahteven problem. Iz tega razloga nam ni uspelo zgraditi odločitvenih modelov na domeni spam za razmerje 1:500 in domenah musk-v2 in spam za razmerje 1:1000. Testiranje smo opravljali na dveh računalnikih — 32-bitni računalnik z dvema dvojedernima procesorjema Intel Core i7 920 pri frekvenci 2.66 GHz in 64-bitni računalnik s štirijedernim procesorjem Intel Xenon X3460 pri frekvenci 2.8 GHz — in kljub temu, da sta računalnika noč in dan 11 dni gradila odločitvene modele, jima ni uspelo dokončati gradnje za omenjene domene in razmerja. Vsekakor bi bilo potrebno za nadaljnje delo uporabiti več računsko zmogljivejših računalnikov.

Odločitvena drevesa bi bilo potrebno zgraditi še na domenah musk-v2 in spam za razmerji 1:500 in 1:1000. Prav tako bi bilo potrebno zgraditi odločitvena drevesa za razmerji razredov 1:5000 in 1:10000, da bi simulirali močno neuravnotežene množice. Zaradi časovne zahtevnosti problema smo vzorčenje realizirali na konstanten, sicer pa ne najbolj optimalen način. Vzorčenje bi lahko spremenili tako, da bi vsakič obdržali vse primere manjšinskega razreda in bi zgolj zametavali ali razmnoževali primere večinskega razreda do željenega razmerja.

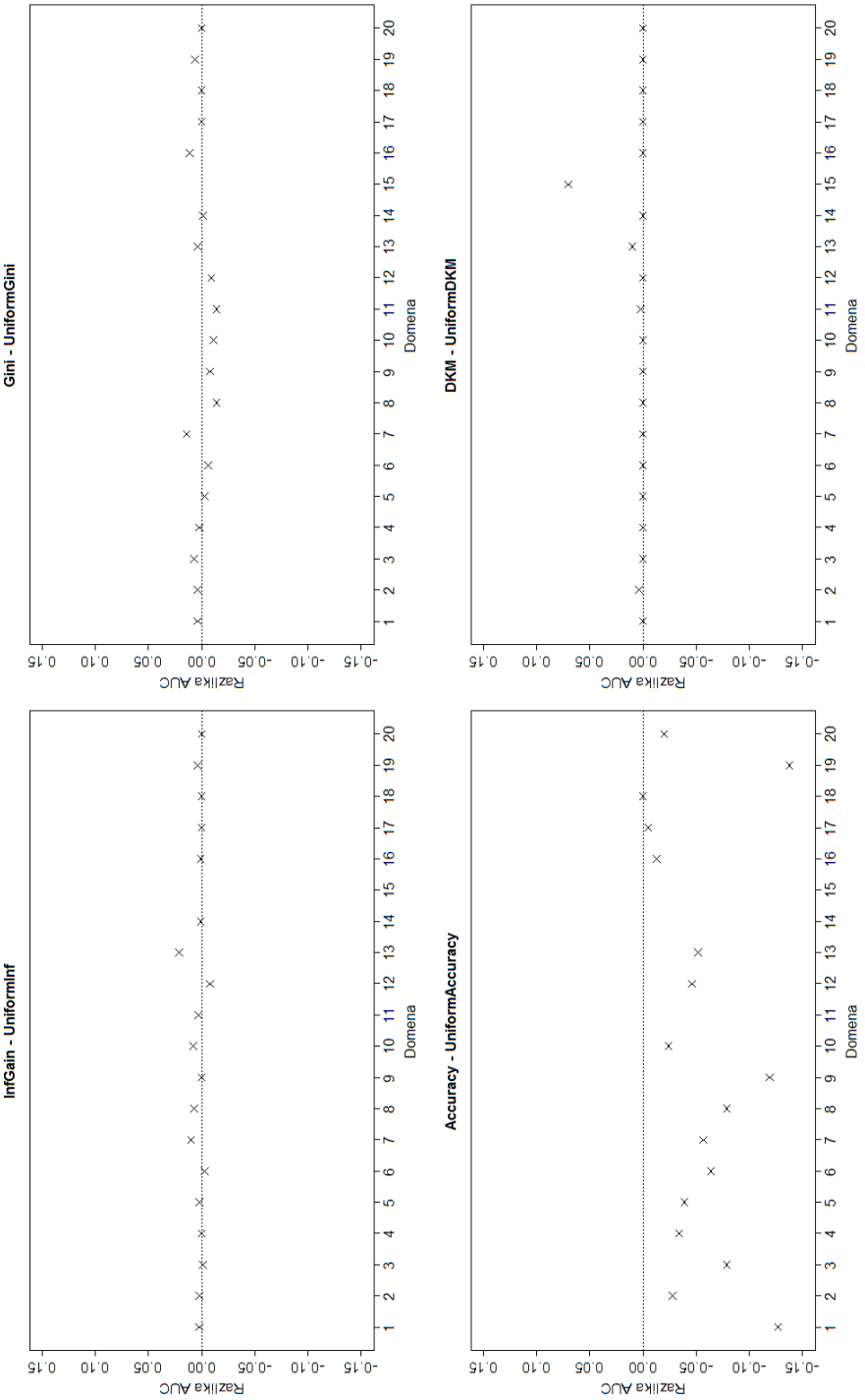
Za vsa razmerja bi bilo potrebno zgraditi še modele naključnih gozdov in primerjati, delovanje mer. Mere bi bilo potrebno testirati tudi na drugih realnih domenah.

Dodatek A

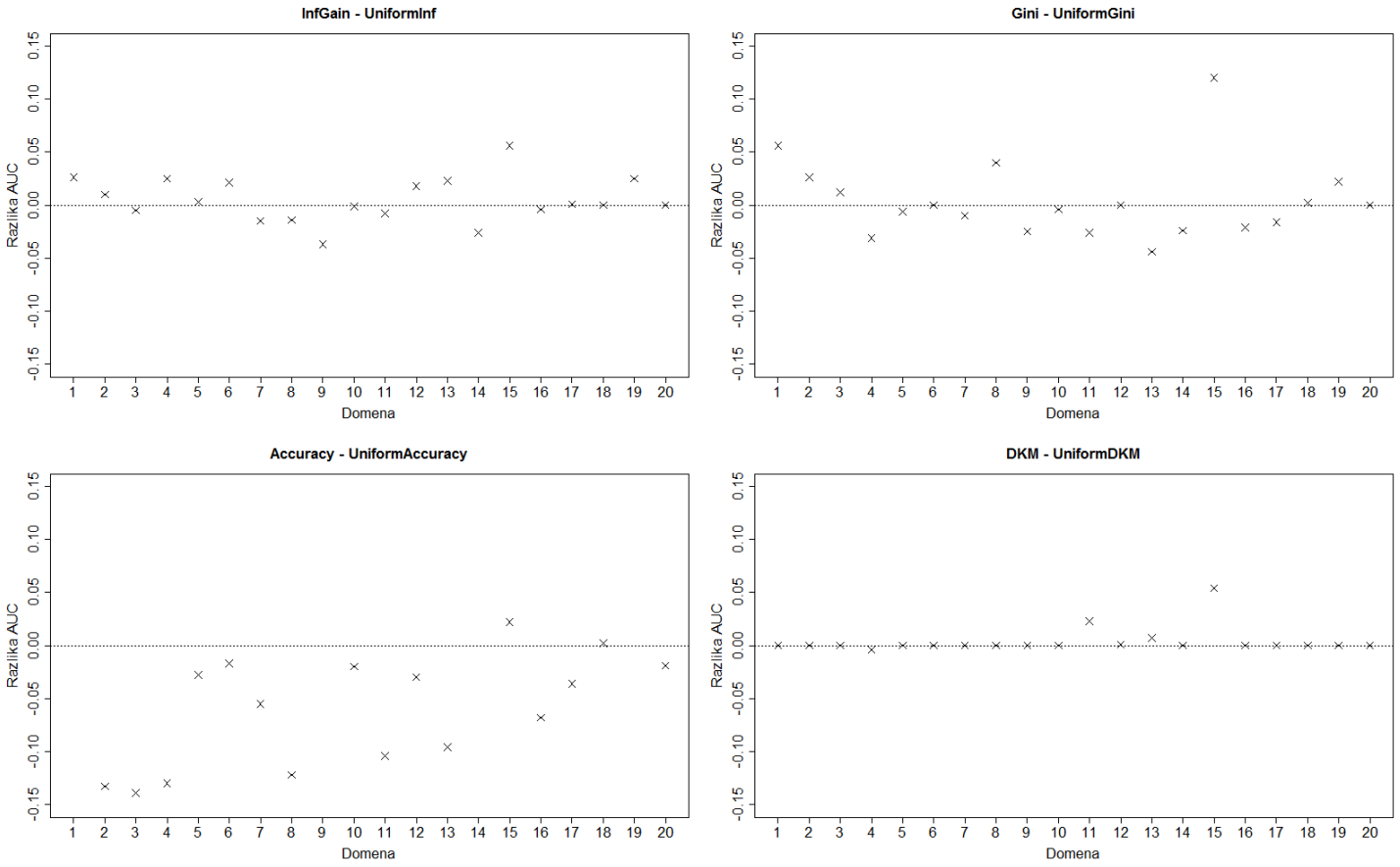
Razlike med klasičnimi in uniform merami

V tem dodatku so s pomočjo grafov prikazane razlike med ocenami AUC mer informacijski prispevek in uniform informacijski prispevek, Gini-indeks in uniform Gini-indeks, klasifikacijska točnost in uniform klasifikacijska točnost ter DKM in uniform DKM.

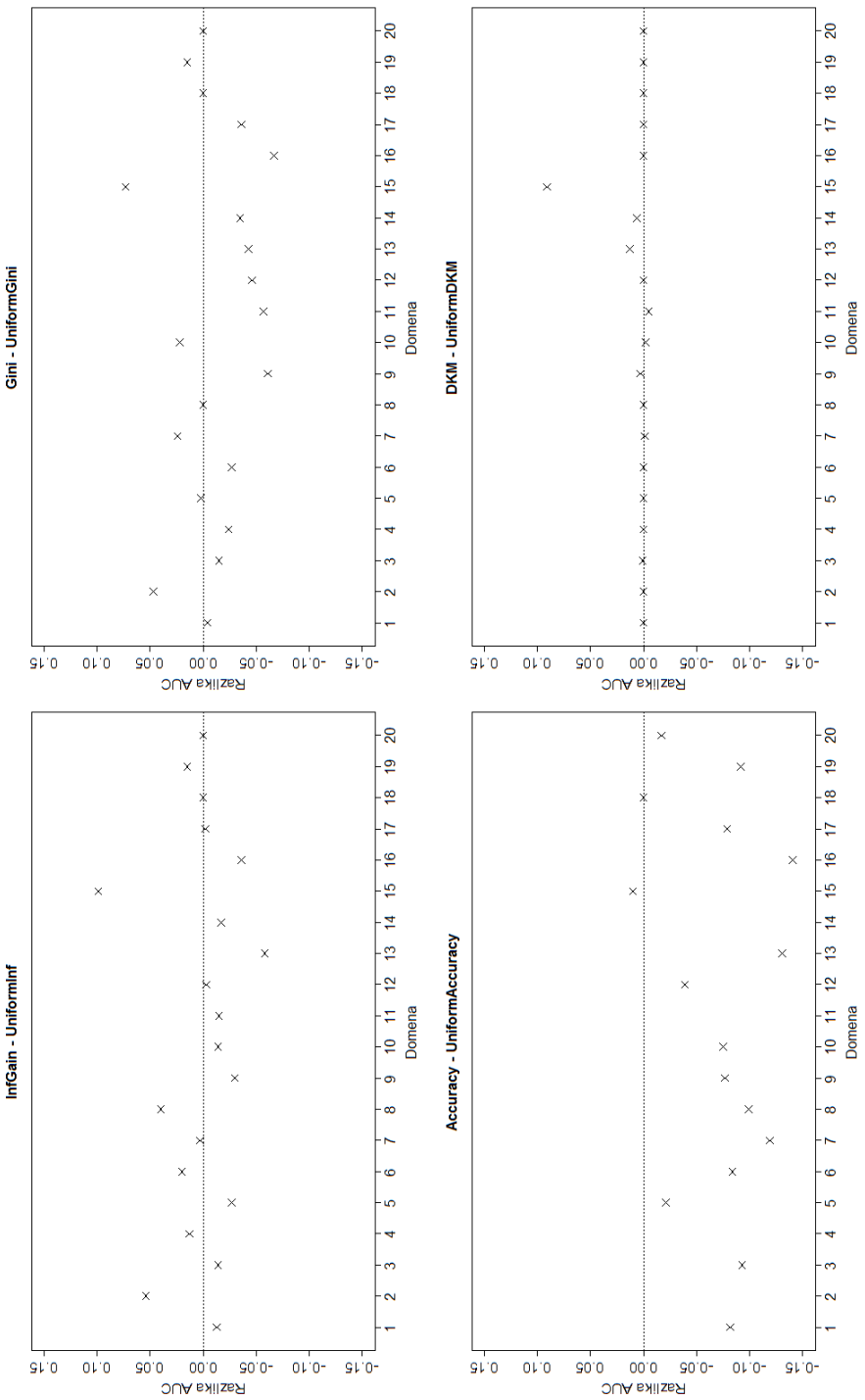
V grafih os y predstavlja razliko med ocenami AUC mer, os x pa zaporedno številko domene (tabela 3.1) na kateri je bil izračunan AUC. Na y osi je prikazana razlika med AUC mer do 15%, vsaka večja razlika na grafu ni prikazana. Pri razmerjih 1:500 in 1:1000 je zaporedje domen enako, z razliko, da se domena, na kateri nismo uspeli zgraditi odločitvenih dreves, ne upošteva in razlika v AUC ni narisana. Nad vsakim grafom je napisan naslov, ki predstavlja meri katerih razlika ocen AUC je narisana. Če v naslovu piše Accuracy − UniformAccuracy, to pomeni, da smo izračunali in narisali razliko v ocenah AUC med klasifikacijsko točnostjo in uniform klasifikacijsko točnostjo.



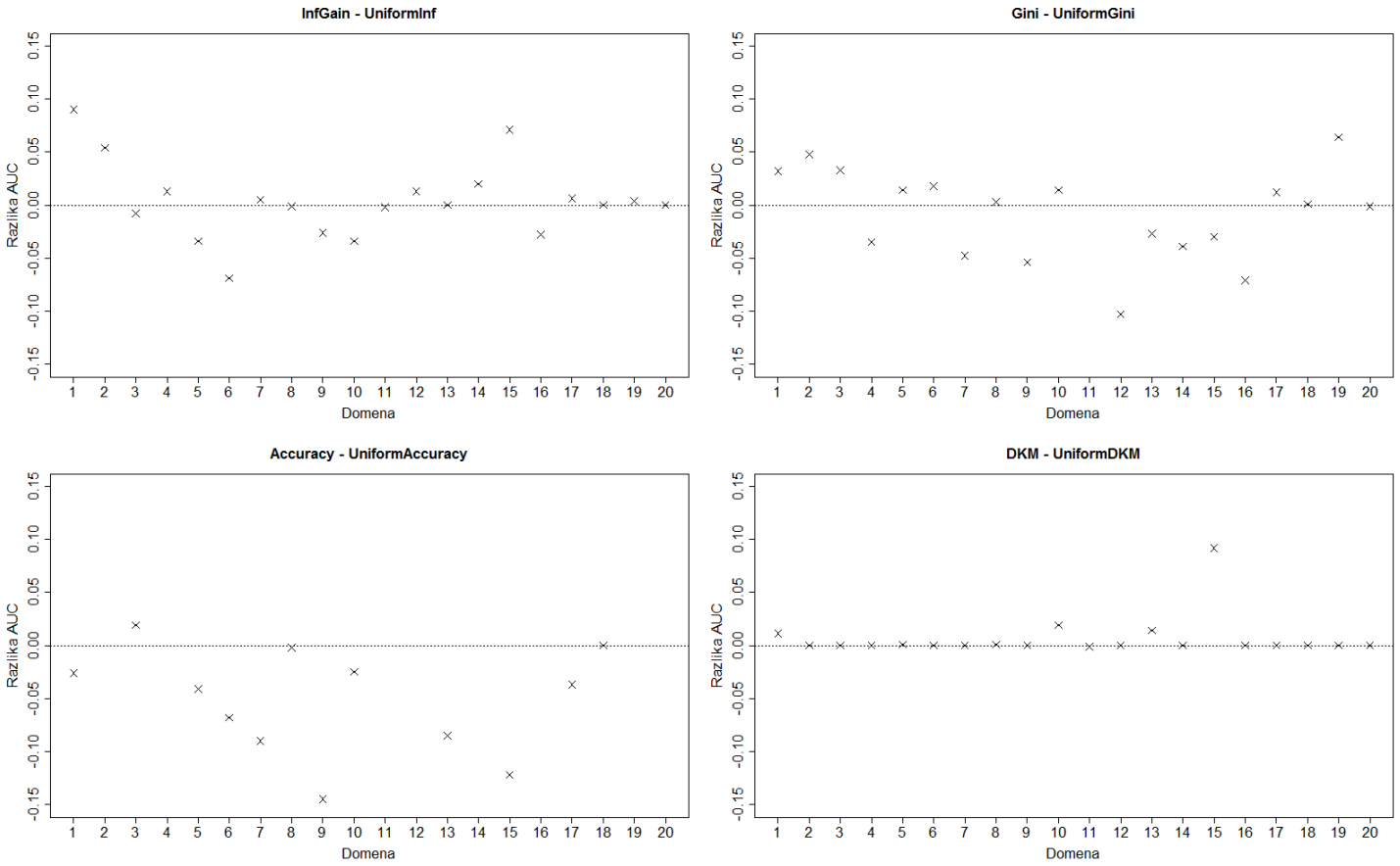
Slika A.1: Primerjava razlik AUC mer pri naravnem razmerju razredov.



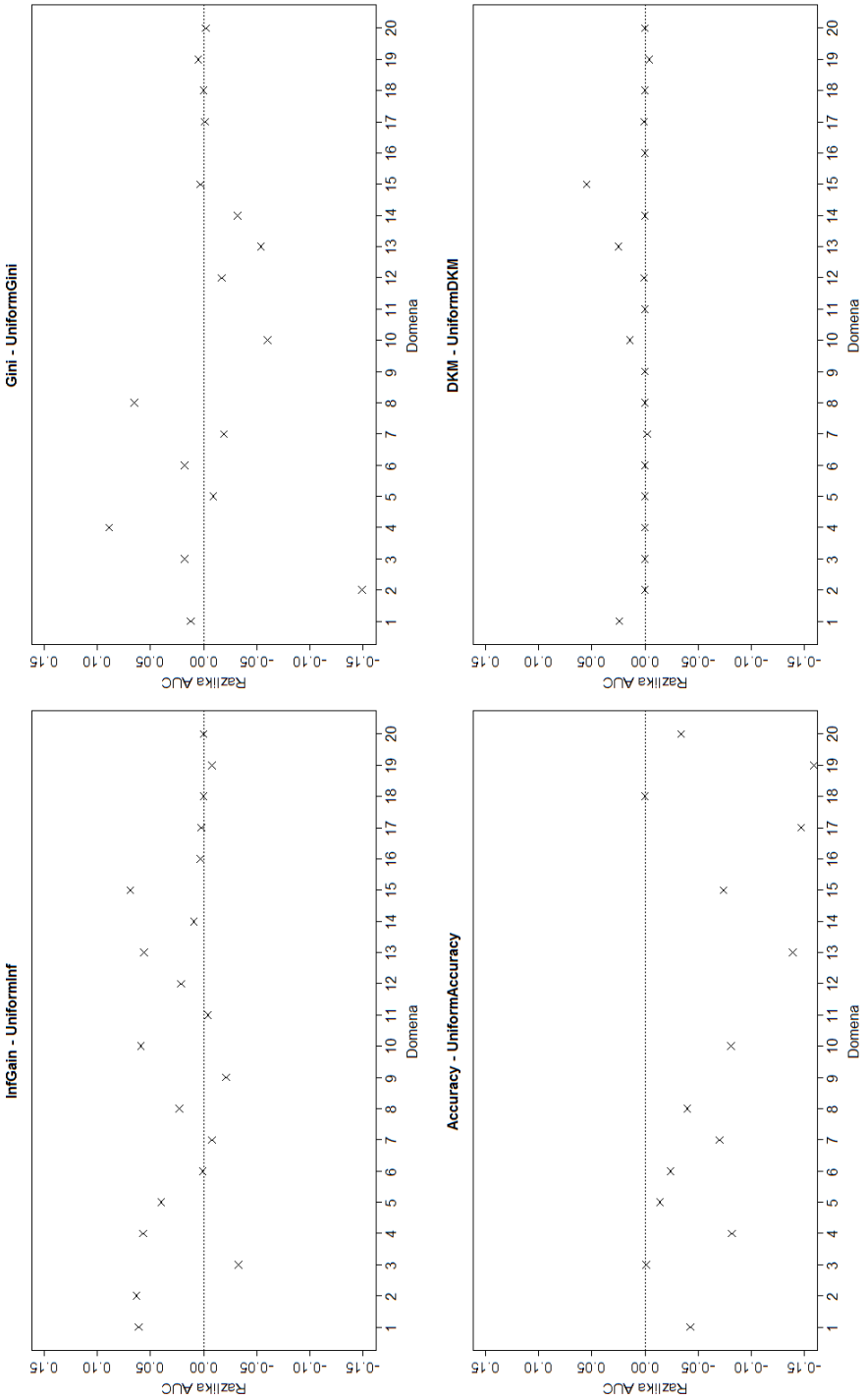
Slika A.2: Primerjava razlik AUC mer pri razmerju razredov 1:5.



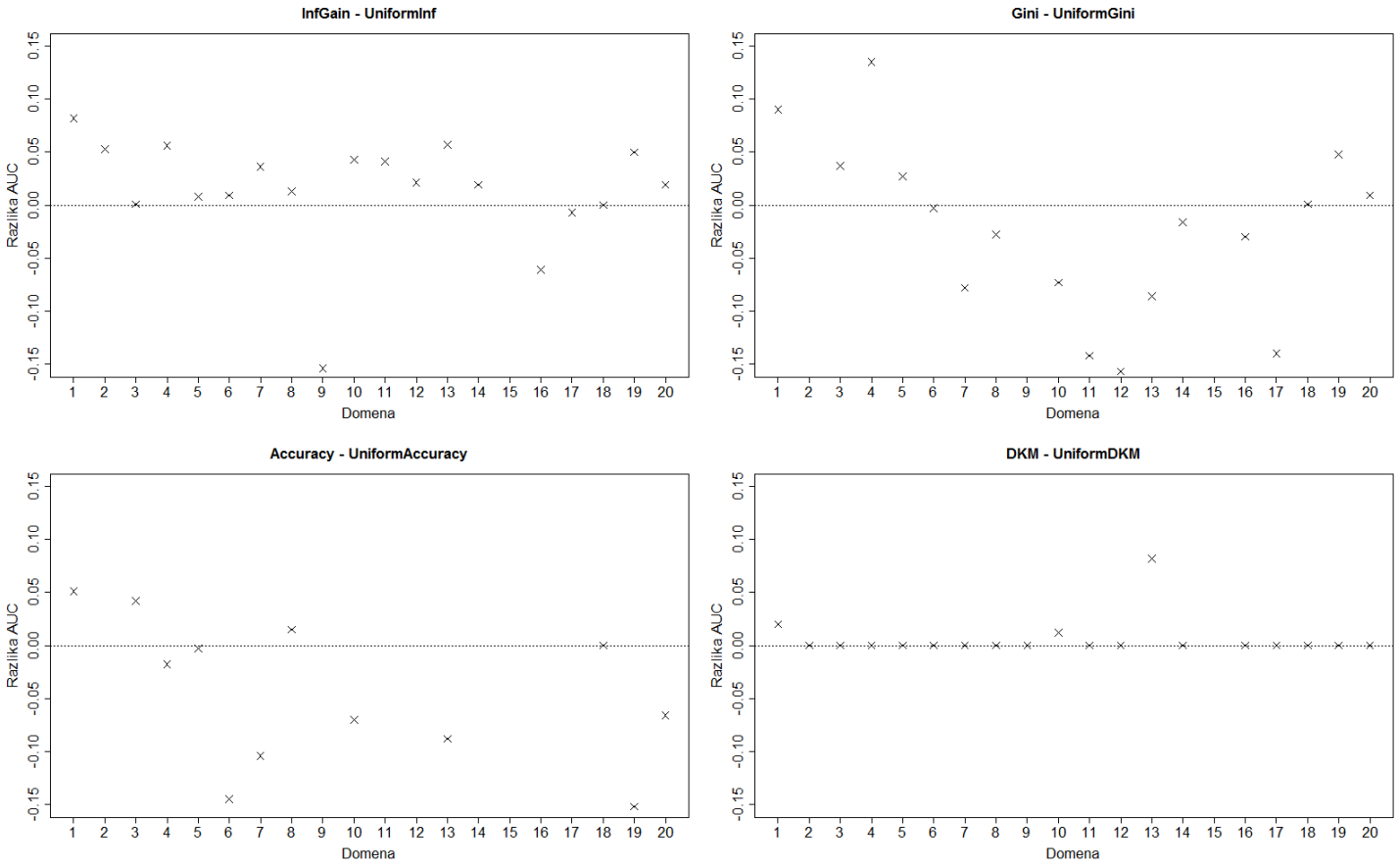
Slika A.3: Primerjava razlik AUC mer pri razmerju razredov 1:10.



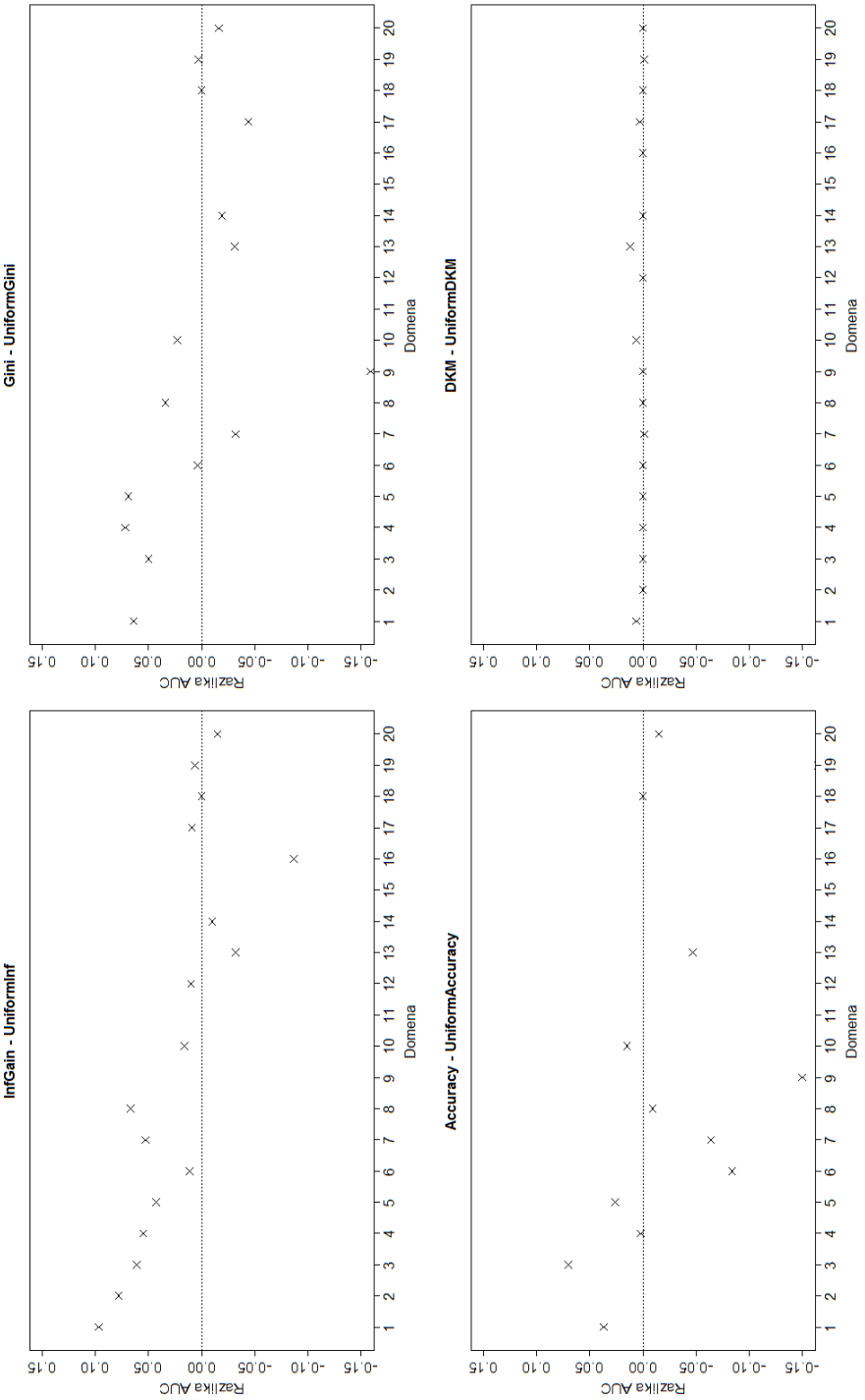
Slika A.4: Primerjava razlik AUC mer pri razmerju razredov 1:50.



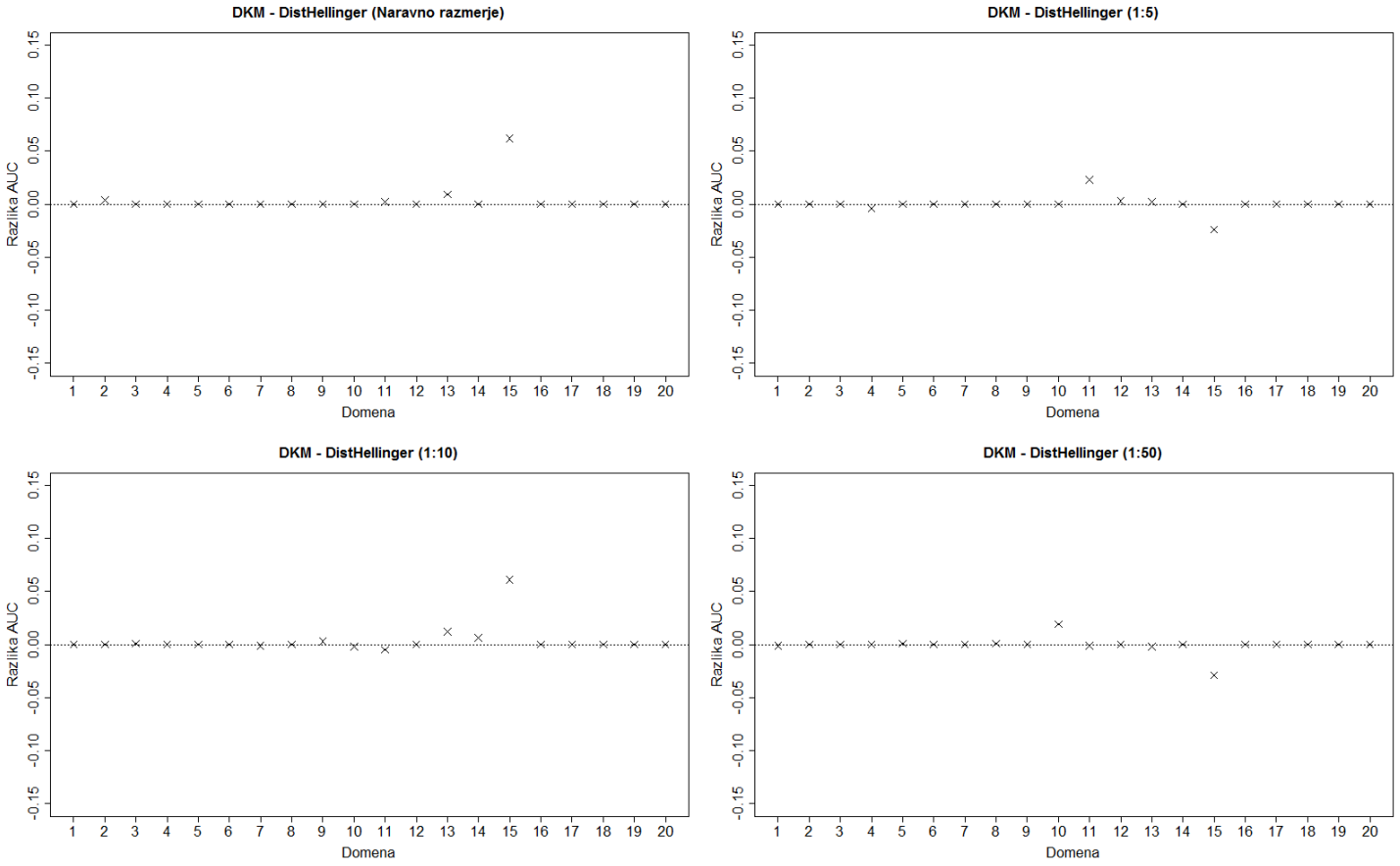
Slika A.5: Primerjava razlik AUC mer pri razmerju razredov 1:100.



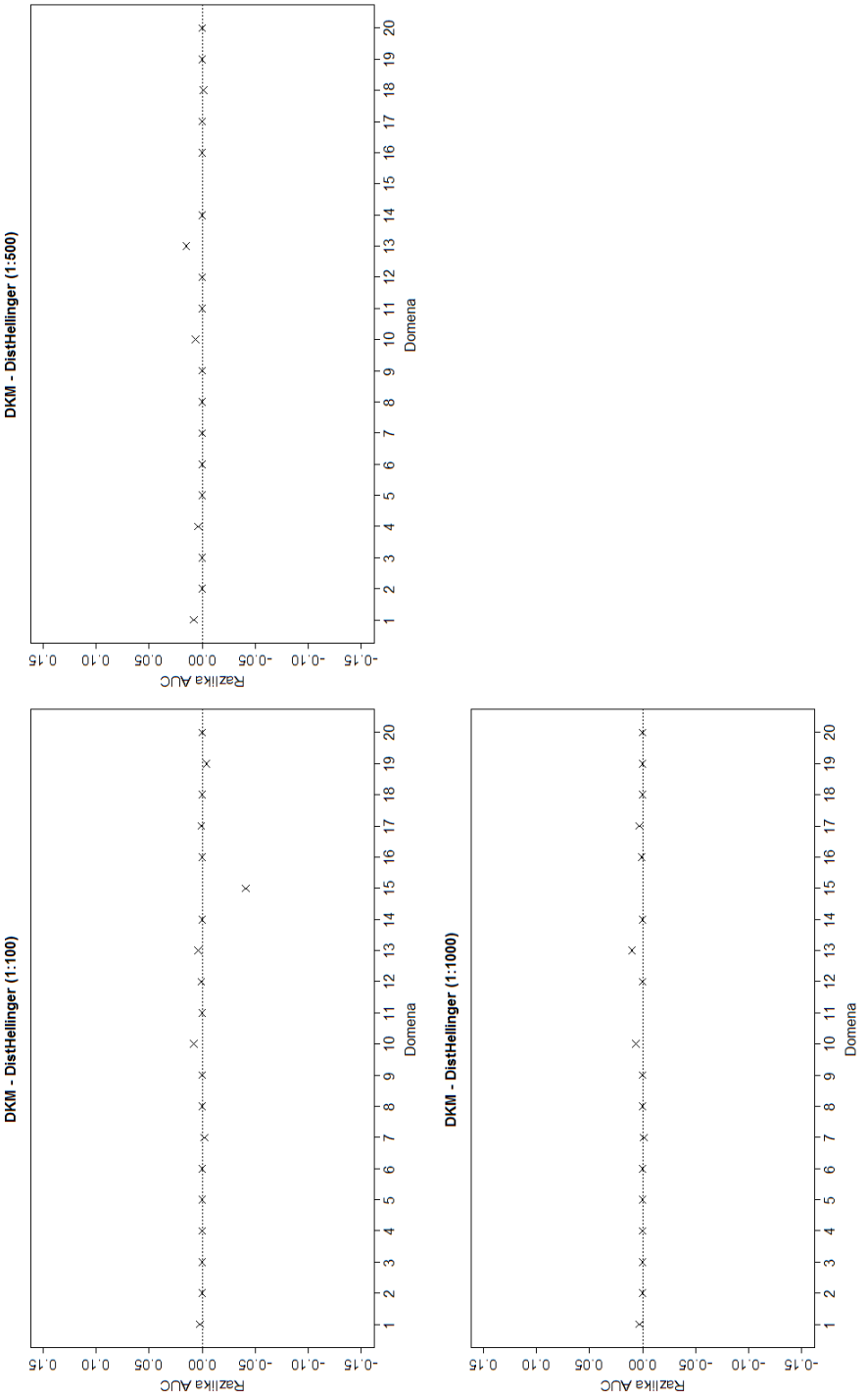
Slika A.6: Primerjava razlik AUC mer pri razmerju razredov 1:500.



Slika A.7: Primerjava razlik AUC mer pri razmerju razredov 1:1000.



Slika A.8: Primerjava razlik AUC mer DKM in Hellingerjeve razdalje (a).



Slika A.9: Primerjava razlik AUC mer DKM in Hellingerjeve razdalje (b).

Dodatek B

AUC odločitvenih dreves

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.871/0.003	0.895/0.003	0.9/0.003	0.892/0.003	0.747/0.056	0.897/0.003	0.897/0.002
australian-credit	0.894/0.023	0.895/0.019	0.903/0.009	0.891/0.017	0.861/0.016	0.898/0.013	0.894/0.012
blood-transfusion	0.718/0.021	0.687/0.03	0.693/0.025	0.696/0.029	0.615/0.083	0.703/0.03	0.703/0.03
cod-rna	0.968/0.002	0.967/0.001	0.969/0.001	0.968/0.003	0.935/0.052	0.965/0.003	0.965/0.003
contraceptive	0.71/0.02	0.72/0.019	0.72/0.03	0.719/0.027	0.684/0.022	0.718/0.022	0.718/0.022
fourclass	0.973/0.013	0.964/0.014	0.967/0.011	0.965/0.009	0.907/0.046	0.973/0.015	0.973/0.015
gamma-telescope	0.866/0.004	0.873/0.004	0.876/0.006	0.877/0.004	0.807/0.012	0.871/0.006	0.871/0.006
german-credit	0.701/0.019	0.701/0.035	0.693/0.02	0.671/0.047	0.608/0.032	0.681/0.034	0.681/0.034
letter-recognition	0.954/0.004	0.944/0.006	0.943/0.004	0.939/0.007	0.827/0.036	0.944/0.005	0.944/0.005
mammographic-mass	0.871/0.01	0.86/0.032	0.862/0.03	0.863/0.032	0.85/0.033	0.842/0.038	0.842/0.038
musk-v2	0.927/0.014	0.929/0.016	0.931/0.017	0.912/0.014	0.666/0.02	0.923/0.017	0.921/0.016
page-blocks	0.93/0.016	0.944/0.02	0.949/0.014	0.94/0.024	0.905/0.017	0.95/0.007	0.95/0.007
pima-diabetes	0.77/0.017	0.773/0.03	0.779/0.021	0.766/0.027	0.704/0.036	0.768/0.03	0.758/0.029
segmentation	0.976/0.006	0.978/0.003	0.979/0.005	0.978/0.006	0.783/0.072	0.98/0.007	0.98/0.007
spam	0.551/0.084	0.746/0.059	0.792/0.027	0.771/0.019	0.717/0.013	0.558/0.056	0.488/0.037
splice	0.967/0.007	0.97/0.009	0.968/0.009	0.973/0.006	0.947/0.009	0.971/0.011	0.971/0.011
svmguide1	0.974/0.007	0.98/0.004	0.981/0.006	0.979/0.005	0.977/0.009	0.98/0.006	0.98/0.006
svmguide3	0.694/0.027	0.5/0	0.742/0.032	0.5/0	0.5/0	0.5/0	0.5/0
tic-tac-toe	0.964/0.013	0.952/0.017	0.951/0.018	0.952/0.015	0.818/0.047	0.951/0.019	0.951/0.019
titanic	0.721/0.023	0.734/0.018	0.725/0.014	0.734/0.018	0.727/0.018	0.733/0.017	0.733/0.017

Tabela B.1: AUC in standardni odklon AUC odločitvenih dreves na domenah z naravnim razmerjem razredov.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.888/0.004	0.893/0.003	0.874/0.004	0.897/0.002	0.874/0.004	0.864/0.004	0.874/0.005
australian-credit	0.887/0.02	0.893/0.019	0.889/0.019	0.894/0.012	0.897/0.015	0.885/0.019	0.893/0.024
blood-transfusion	0.689/0.029	0.688/0.029	0.694/0.028	0.703/0.03	0.694/0.028	0.694/0.03	0.694/0.028
cod-rna	0.966/0.001	0.967/0.002	0.969/0.002	0.965/0.003	0.969/0.002	0.969/0.002	0.969/0.002
contraceptive	0.722/0.022	0.718/0.016	0.723/0.017	0.718/0.022	0.724/0.017	0.714/0.018	0.72/0.019
fourclass	0.971/0.013	0.967/0.016	0.971/0.011	0.973/0.015	0.971/0.011	0.974/0.009	0.971/0.011
gamma-telescope	0.863/0.006	0.863/0.008	0.864/0.007	0.871/0.006	0.864/0.007	0.87/0.004	0.864/0.007
german-credit	0.685/0.024	0.694/0.033	0.687/0.027	0.681/0.034	0.683/0.024	0.68/0.02	0.678/0.028
letter-recognition	0.947/0.003	0.944/0.007	0.946/0.005	0.944/0.005	0.947/0.004	0.944/0.007	0.947/0.004
mammographic-mass	0.874/0.017	0.852/0.034	0.874/0.012	0.842/0.038	0.875/0.015	0.871/0.018	0.876/0.016
musk-v2	0.926/0.013	0.926/0.013	0.92/0.014	0.921/0.016	0.92/0.013	0.926/0.008	0.92/0.013
page-blocks	0.949/0.012	0.952/0.013	0.951/0.015	0.95/0.007	0.95/0.015	0.942/0.019	0.95/0.015
pima-diabetes	0.762/0.023	0.752/0.035	0.756/0.027	0.759/0.029	0.756/0.027	0.764/0.033	0.756/0.027
segmentation	0.979/0.007	0.977/0.008	0.978/0.005	0.98/0.007	0.978/0.005	0.98/0.005	0.978/0.005
spam	0.488/0.037	0.488/0.037	0.472/0.077	0.496/0.059	0.473/0.078	0.502/0.084	0.473/0.078
splice	0.962/0.01	0.969/0.007	0.96/0.008	0.971/0.011	0.952/0.009	0.945/0.014	0.961/0.009
svmguide1	0.979/0.007	0.98/0.005	0.982/0.005	0.98/0.006	0.982/0.005	0.981/0.006	0.982/0.005
svmguide3	0.5/0	0.5/0	0.5/0	0.5/0	0.5/0	0.5/0	0.692/0.031
tic-tac-toe	0.946/0.016	0.948/0.019	0.956/0.014	0.951/0.019	0.95/0.016	0.952/0.016	0.955/0.014
titanic	0.734/0.018	0.734/0.018	0.747/0.013	0.733/0.017	0.747/0.013	0.749/0.012	0.738/0.023

Tabela B.2: Nadaljevanje tabele B.1.

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.767/0.031	0.751/0.017	0.743/0.046	0.721/0.071	0.571/0.01	0.739/0.008	0.739/0.008
australian-credit	0.874/0.035	0.902/0.012	0.895/0.025	0.908/0.018	0.733/0.064	0.873/0.042	0.873/0.042
blood-transfusion	0.677/0.061	0.657/0.087	0.646/0.068	0.667/0.073	0.527/0.038	0.683/0.024	0.683/0.024
cod-rna	0.855/0.045	0.864/0.04	0.873/0.032	0.827/0.074	0.729/0.088	0.843/0.044	0.847/0.036
contraceptive	0.59/0.006	0.63/0.055	0.61/0.005	0.621/0.078	0.608/0.055	0.596/0.029	0.596/0.029
fourclass	0.895/0.062	0.921/0.031	0.912/0.03	0.896/0.079	0.884/0.049	0.91/0.061	0.91/0.061
gamma-telescope	0.764/0.042	0.741/0.047	0.736/0.064	0.696/0.054	0.66/0.07	0.772/0.074	0.772/0.074
german-credit	0.669/0.063	0.634/0.063	0.673/0.021	0.667/0.022	0.541/0.054	0.65/0.027	0.65/0.027
letter-recognition	0.696/0.071	0.61/0.098	0.616/0.103	0.632/0.123	0.52/0.021	0.618/0.135	0.618/0.135
mammographic-mass	0.865/0.021	0.854/0.025	0.861/0.029	0.853/0.018	0.849/0.017	0.849/0.009	0.849/0.009
musk-v2	0.778/0.052	0.796/0.034	0.785/0.041	0.761/0.053	0.67/0.059	0.774/0.042	0.751/0.06
page-blocks	0.878/0.02	0.903/0.051	0.898/0.05	0.896/0.055	0.874/0.051	0.892/0.034	0.891/0.034
pima-diabetes	0.71/0.033	0.744/0.033	0.75/0.047	0.709/0.048	0.633/0.062	0.733/0.038	0.726/0.056
segmentation	0.934/0.025	0.913/0.031	0.922/0.031	0.912/0.032	0.656/0.045	0.939/0.027	0.939/0.027
spam	0.636/0.108	0.563/0.061	0.603/0.066	0.627/0.089	0.604/0.113	0.561/0.033	0.507/0.023
splice	0.886/0.037	0.919/0.028	0.892/0.031	0.902/0.038	0.875/0.026	0.928/0.036	0.928/0.036
svmguide1	0.952/0.042	0.957/0.032	0.942/0.036	0.94/0.034	0.92/0.027	0.953/0.029	0.953/0.029
svmguide3	0.68/0.038	0.499/0.002	0.641/0.049	0.501/0.004	0.501/0.004	0.499/0.002	0.499/0.002
tic-tac-toe	0.863/0.037	0.803/0.069	0.836/0.042	0.814/0.054	0.636/0.052	0.778/0.044	0.778/0.044
titanic	0.705/0.01	0.707/0.018	0.706/0.01	0.707/0.018	0.69/0.024	0.707/0.019	0.707/0.019

Tabela B.3: AUC in standardni odklon AUC odločitvenih dreves na domenah z razmerjem razredov 1:5.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.665/0.047	0.725/0.011	0.751/0.053	0.739/0.008	0.762/0.054	0.734/0.021	0.76/0.024
australian-credit	0.882/0.052	0.892/0.03	0.866/0.066	0.873/0.042	0.893/0.018	0.855/0.066	0.833/0.102
blood-transfusion	0.655/0.083	0.662/0.06	0.666/0.085	0.683/0.024	0.666/0.085	0.664/0.075	0.666/0.085
cod-rna	0.858/0.045	0.839/0.041	0.859/0.04	0.847/0.036	0.859/0.04	0.835/0.025	0.859/0.04
contraceptive	0.627/0.064	0.627/0.064	0.636/0.012	0.596/0.029	0.657/0.063	0.666/0.057	0.592/0.039
fourclass	0.896/0.053	0.9/0.057	0.901/0.062	0.91/0.061	0.901/0.062	0.882/0.073	0.901/0.062
gamma-telescope	0.706/0.066	0.756/0.059	0.715/0.035	0.772/0.074	0.715/0.035	0.746/0.058	0.715/0.035
german-credit	0.627/0.081	0.648/0.063	0.663/0.018	0.65/0.027	0.681/0.039	0.706/0.032	0.68/0.035
letter-recognition	0.657/0.048	0.647/0.134	0.693/0.024	0.618/0.135	0.693/0.024	0.726/0.071	0.693/0.024
mammographic-mass	0.857/0.023	0.855/0.021	0.869/0.025	0.849/0.009	0.866/0.023	0.875/0.024	0.864/0.029
musk-v2	0.787/0.065	0.804/0.045	0.774/0.061	0.751/0.06	0.8/0.046	0.779/0.044	0.8/0.046
page-blocks	0.896/0.036	0.885/0.036	0.904/0.027	0.889/0.032	0.904/0.027	0.898/0.045	0.904/0.027
pima-diabetes	0.753/0.05	0.721/0.057	0.729/0.036	0.731/0.053	0.729/0.036	0.731/0.048	0.729/0.036
segmentation	0.936/0.024	0.939/0.023	0.931/0.025	0.939/0.027	0.931/0.025	0.926/0.028	0.931/0.025
spam	0.507/0.023	0.507/0.023	0.582/0.061	0.585/0.062	0.582/0.061	0.582/0.061	0.582/0.061
splice	0.923/0.037	0.923/0.037	0.943/0.029	0.928/0.036	0.957/0.029	0.844/0.08	0.927/0.051
svmguide1	0.956/0.032	0.956/0.032	0.956/0.032	0.953/0.029	0.956/0.032	0.956/0.032	0.956/0.032
svmguide3	0.499/0.002	0.499/0.002	0.499/0.002	0.499/0.002	0.5/0.004	0.499/0.002	0.664/0.07
tic-tac-toe	0.792/0.053	0.778/0.044	0.817/0.047	0.778/0.044	0.831/0.076	0.834/0.043	0.838/0.05
titanic	0.707/0.018	0.707/0.019	0.709/0.019	0.707/0.019	0.709/0.019	0.709/0.019	0.709/0.019

Tabela B.4: Nadaljevanje tabele B.3.

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.822/0.052	0.796/0.053	0.815/0.069	0.802/0.049	0.686/0.114	0.817/0.042	0.817/0.042
australian-credit	0.86/0.019	0.853/0.071	0.896/0.019	0.885/0.051	0.645/0.08	0.801/0.098	0.801/0.098
blood-transfusion	0.73/0.018	0.707/0.033	0.698/0.042	0.684/0.04	0.606/0.1	0.715/0.027	0.714/0.026
cod-rna	0.85/0.055	0.848/0.051	0.846/0.046	0.808/0.087	0.696/0.082	0.842/0.051	0.842/0.051
contraceptive	0.546/0.021	0.597/0.045	0.619/0.029	0.61/0.044	0.62/0.059	0.584/0.078	0.584/0.078
fourclass	0.914/0.06	0.934/0.032	0.9/0.089	0.882/0.119	0.836/0.08	0.958/0.015	0.958/0.015
gamma-telescope	0.759/0.061	0.712/0.064	0.717/0.073	0.723/0.042	0.589/0.071	0.73/0.084	0.731/0.083
german-credit	0.639/0.045	0.648/0.044	0.601/0.061	0.611/0.058	0.544/0.047	0.624/0.037	0.624/0.037
letter-recognition	0.72/0.025	0.733/0.084	0.693/0.087	0.706/0.084	0.638/0.07	0.725/0.051	0.722/0.05
mammographic-mass	0.852/0.053	0.848/0.02	0.853/0.072	0.873/0.036	0.791/0.078	0.845/0.039	0.847/0.035
musk-v2	0.817/0.075	0.796/0.06	0.782/0.041	0.744/0.081	0.617/0.059	0.784/0.026	0.789/0.034
page-blocks	0.865/0.029	0.878/0.044	0.889/0.047	0.856/0.064	0.864/0.054	0.887/0.038	0.887/0.038
pima-diabetes	0.736/0.06	0.673/0.046	0.698/0.044	0.686/0.053	0.6/0.075	0.732/0.042	0.719/0.056
segmentation	0.939/0.016	0.924/0.025	0.926/0.019	0.903/0.021	0.758/0.084	0.925/0.045	0.919/0.051
spam	0.654/0.079	0.599/0.084	0.572/0.089	0.573/0.068	0.547/0.088	0.591/0.061	0.5/0
splice	0.824/0.05	0.917/0.037	0.858/0.06	0.876/0.071	0.788/0.052	0.92/0.016	0.92/0.016
svmguide1	0.94/0.038	0.966/0.025	0.959/0.019	0.921/0.043	0.875/0.097	0.948/0.031	0.948/0.031
svmguide3	0.644/0.053	0.5/0.003	0.625/0.087	0.5/0	0.5/0	0.5/0.003	0.5/0.003
tic-tac-toe	0.794/0.067	0.754/0.049	0.772/0.031	0.764/0.033	0.657/0.027	0.725/0.03	0.725/0.03
titanic	0.721/0.036	0.718/0.039	0.721/0.036	0.718/0.039	0.704/0.046	0.718/0.039	0.718/0.039

Tabela B.5: AUC in standardni odklon AUC odločitvenih dreves na domenah z razmerjem razredov 1:10.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.806/0.043	0.809/0.039	0.768/0.044	0.817/0.042	0.794/0.054	0.804/0.034	0.768/0.073
australian-credit	0.838/0.103	0.799/0.117	0.828/0.085	0.801/0.098	0.873/0.036	0.797/0.066	0.811/0.043
blood-transfusion	0.699/0.014	0.721/0.037	0.699/0.017	0.714/0.026	0.696/0.016	0.685/0.028	0.696/0.016
cod-rna	0.832/0.05	0.835/0.047	0.862/0.034	0.842/0.051	0.873/0.036	0.863/0.058	0.873/0.036
contraceptive	0.608/0.068	0.624/0.052	0.641/0.051	0.584/0.078	0.621/0.037	0.632/0.03	0.647/0.062
fourclass	0.909/0.061	0.914/0.062	0.92/0.067	0.958/0.015	0.92/0.067	0.911/0.084	0.92/0.067
gamma-telescope	0.699/0.075	0.709/0.067	0.708/0.042	0.731/0.083	0.708/0.042	0.721/0.036	0.708/0.042
german-credit	0.611/0.048	0.608/0.054	0.643/0.037	0.624/0.037	0.635/0.035	0.664/0.06	0.633/0.066
letter-recognition	0.767/0.017	0.763/0.029	0.715/0.059	0.722/0.05	0.715/0.059	0.709/0.04	0.715/0.059
mammographic-mass	0.851/0.025	0.862/0.032	0.866/0.028	0.847/0.035	0.858/0.04	0.867/0.03	0.869/0.028
musk-v2	0.801/0.048	0.811/0.049	0.807/0.053	0.789/0.034	0.807/0.053	0.803/0.042	0.807/0.053
page-blocks	0.902/0.026	0.881/0.039	0.903/0.048	0.887/0.038	0.903/0.048	0.903/0.039	0.903/0.048
pima-diabetes	0.729/0.041	0.731/0.063	0.731/0.051	0.72/0.053	0.732/0.053	0.736/0.05	0.732/0.053
segmentation	0.938/0.021	0.941/0.017	0.944/0.01	0.919/0.051	0.939/0.021	0.929/0.035	0.939/0.021
spam	0.5/0	0.5/0	0.537/0.09	0.53/0.068	0.537/0.09	0.538/0.092	0.537/0.09
splice	0.943/0.031	0.953/0.017	0.929/0.02	0.92/0.016	0.931/0.015	0.869/0.068	0.886/0.075
svmguide1	0.957/0.023	0.968/0.016	0.954/0.027	0.948/0.031	0.954/0.027	0.951/0.013	0.954/0.027
svmguide3	0.5/0.003	0.5/0.003	0.5/0.003	0.5/0.003	0.5/0.003	0.5/0.003	0.66/0.052
tic-tac-toe	0.749/0.018	0.739/0.026	0.749/0.034	0.725/0.03	0.73/0.06	0.728/0.035	0.734/0.04
titanic	0.718/0.039	0.718/0.039	0.721/0.039	0.718/0.039	0.721/0.039	0.721/0.039	0.721/0.039

Tabela B.6: Nadaljevanje tabele B.5.

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.731/0.076	0.724/0.042	0.695/0.134	0.69/0.082	0.653/0.073	0.742/0.032	0.731/0.032
australian-credit	0.861/0.04	0.883/0.025	0.896/0.027	0.869/0.05	0.601/0.097	0.854/0.034	0.854/0.034
blood-transfusion	0.603/0.063	0.627/0.018	0.601/0.048	0.641/0.029	0.638/0.046	0.631/0.024	0.631/0.024
cod-rna	0.812/0.049	0.804/0.045	0.804/0.088	0.775/0.072	0.638/0.101	0.779/0.061	0.779/0.061
contraceptive	0.607/0.13	0.601/0.091	0.594/0.03	0.624/0.03	0.602/0.063	0.592/0.083	0.591/0.081
fourclass	0.911/0.044	0.84/0.052	0.855/0.051	0.922/0.043	0.867/0.039	0.931/0.007	0.931/0.007
gamma-telescope	0.745/0.078	0.707/0.091	0.707/0.091	0.662/0.106	0.613/0.082	0.7/0.043	0.7/0.043
german-credit	0.61/0.052	0.601/0.05	0.615/0.034	0.608/0.047	0.568/0.043	0.625/0.055	0.624/0.056
letter-recognition	0.778/0.015	0.697/0.017	0.667/0.017	0.672/0.064	0.545/0.054	0.699/0.022	0.699/0.022
mammographic-mass	0.855/0.028	0.832/0.036	0.815/0.034	0.855/0.032	0.796/0.033	0.871/0.032	0.852/0.048
musk-v2	0.771/0.037	0.747/0.074	0.762/0.073	0.594/0.074	0.524/0.029	0.755/0.032	0.756/0.033
page-blocks	0.875/0.044	0.872/0.054	0.879/0.046	0.765/0.109	0.645/0.069	0.865/0.051	0.865/0.051
pima-diabetes	0.685/0.037	0.736/0.055	0.725/0.064	0.696/0.077	0.636/0.082	0.742/0.049	0.728/0.035
segmentation	0.913/0.028	0.938/0.043	0.927/0.037	0.88/0.066	0.654/0.078	0.937/0.039	0.937/0.039
spam	0.519/0.152	0.571/0.105	0.55/0.145	0.47/0.098	0.463/0.102	0.544/0.044	0.452/0.077
splice	0.802/0.03	0.873/0.047	0.86/0.067	0.816/0.043	0.622/0.117	0.904/0.018	0.904/0.018
svmguide1	0.937/0.041	0.955/0.031	0.978/0.015	0.958/0.035	0.913/0.06	0.953/0.033	0.953/0.033
svmguide3	0.608/0.031	0.5/0.001	0.579/0.09	0.501/0.003	0.501/0.003	0.5/0.001	0.5/0
tic-tac-toe	0.859/0.006	0.757/0.036	0.771/0.031	0.826/0.05	0.621/0.09	0.773/0.031	0.773/0.031
titanic	0.713/0.021	0.713/0.025	0.714/0.021	0.712/0.024	0.542/0.089	0.713/0.025	0.713/0.025

Tabela B.7: AUC in standardni odklon AUC odločitvenih dreves na domenah z razmerjem razredov 1:50.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.658/0.097	0.634/0.093	0.679/0.059	0.743/0.03	0.701/0.058	0.679/0.083	0.665/0.055
australian-credit	0.821/0.061	0.829/0.083	0.806/0.053	0.854/0.034	0.821/0.043	0.824/0.074	0.824/0.066
blood-transfusion	0.608/0.029	0.635/0.023	0.619/0.041	0.631/0.024	0.618/0.04	0.599/0.039	0.618/0.04
cod-rna	0.81/0.042	0.791/0.075	0.809/0.051	0.779/0.061	0.809/0.051	0.802/0.04	0.809/0.051
contraceptive	0.61/0.042	0.635/0.059	0.643/0.051	0.591/0.081	0.628/0.027	0.593/0.025	0.656/0.043
fourclass	0.904/0.055	0.909/0.056	0.935/0.021	0.931/0.007	0.935/0.021	0.922/0.035	0.935/0.021
gamma-telescope	0.71/0.048	0.702/0.067	0.703/0.055	0.7/0.043	0.703/0.055	0.704/0.05	0.703/0.055
german-credit	0.605/0.048	0.602/0.072	0.57/0.056	0.624/0.056	0.598/0.049	0.575/0.051	0.556/0.059
letter-recognition	0.726/0.059	0.723/0.016	0.69/0.032	0.699/0.022	0.69/0.032	0.67/0.069	0.69/0.032
mammographic-mass	0.841/0.04	0.866/0.031	0.821/0.018	0.852/0.048	0.828/0.004	0.809/0.022	0.797/0.02
musk-v2	0.782/0.049	0.749/0.048	0.77/0.06	0.756/0.033	0.763/0.06	0.764/0.058	0.763/0.06
page-blocks	0.868/0.067	0.859/0.071	0.898/0.044	0.865/0.051	0.898/0.044	0.881/0.043	0.898/0.044
pima-diabetes	0.723/0.055	0.736/0.049	0.721/0.037	0.744/0.036	0.721/0.037	0.705/0.057	0.721/0.037
segmentation	0.919/0.036	0.918/0.04	0.932/0.033	0.937/0.039	0.932/0.033	0.941/0.031	0.932/0.033
spam	0.5/0	0.5/0	0.585/0.106	0.573/0.08	0.585/0.106	0.6/0.121	0.585/0.106
splice	0.887/0.03	0.901/0.06	0.855/0.044	0.904/0.018	0.859/0.042	0.863/0.022	0.899/0.016
svmguide1	0.946/0.034	0.949/0.031	0.95/0.035	0.953/0.033	0.95/0.035	0.954/0.032	0.95/0.035
svmguide3	0.5/0.001	0.5/0.001	0.501/0.003	0.5/0.001	0.5/0	0.5/0.001	0.613/0.034
tic-tac-toe	0.762/0.036	0.753/0.038	0.811/0.063	0.773/0.031	0.772/0.027	0.801/0.029	0.788/0.04
titanic	0.713/0.025	0.713/0.025	0.717/0.029	0.713/0.025	0.715/0.028	0.717/0.029	0.717/0.029

Tabela B.8: Nadaljevanje tabele B.7.

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.692/0.087	0.674/0.079	0.738/0.046	0.665/0.093	0.603/0.072	0.655/0.049	0.631/0.074
australian-credit	0.857/0.025	0.903/0.014	0.892/0.031	0.66/0.179	0.551/0.039	0.848/0.011	0.848/0.011
blood-transfusion	0.652/0.026	0.607/0.027	0.603/0.034	0.641/0.04	0.611/0.045	0.625/0.031	0.625/0.031
cod-rna	0.781/0.045	0.791/0.044	0.738/0.1	0.796/0.098	0.657/0.065	0.742/0.056	0.742/0.056
contraceptive	0.527/0.053	0.618/0.056	0.64/0.056	0.614/0.04	0.578/0.049	0.59/0.043	0.59/0.043
fourclass	0.945/0.035	0.899/0.041	0.902/0.025	0.912/0.048	0.883/0.061	0.915/0.013	0.915/0.013
gamma-telescope	0.681/0.07	0.655/0.076	0.692/0.09	0.624/0.1	0.564/0.059	0.653/0.05	0.655/0.051
german-credit	0.625/0.03	0.641/0.044	0.615/0.037	0.637/0.036	0.538/0.034	0.619/0.049	0.619/0.049
letter-recognition	0.778/0.071	0.663/0.013	0.513/0.041	0.517/0.055	0.503/0.009	0.727/0.057	0.727/0.057
mammographic-mass	0.822/0.036	0.855/0.046	0.814/0.046	0.764/0.15	0.696/0.156	0.819/0.042	0.805/0.031
musk-v2	0.778/0.055	0.73/0.072	0.748/0.088	0.566/0.06	0.573/0.079	0.726/0.041	0.726/0.041
page-blocks	0.851/0.057	0.89/0.05	0.917/0.025	0.859/0.098	0.603/0.042	0.871/0.035	0.87/0.035
pima-diabetes	0.682/0.036	0.719/0.035	0.725/0.051	0.635/0.118	0.543/0.07	0.725/0.042	0.7/0.036
segmentation	0.897/0.03	0.926/0.03	0.9/0.036	0.88/0.093	0.698/0.06	0.923/0.029	0.923/0.029
spam	0.574/0.165	0.546/0.103	0.551/0.119	0.48/0.085	0.512/0.073	0.533/0.093	0.478/0.038
splice	0.856/0.036	0.909/0.024	0.866/0.044	0.674/0.035	0.653/0.085	0.896/0.017	0.896/0.017
svmguide1	0.905/0.026	0.936/0.021	0.94/0.024	0.933/0.046	0.787/0.201	0.931/0.027	0.93/0.026
svmguide3	0.586/0.06	0.499/0.001	0.549/0.081	0.499/0.001	0.5/0.004	0.499/0.001	0.499/0.001
tic-tac-toe	0.839/0.034	0.792/0.041	0.76/0.012	0.813/0.038	0.629/0.068	0.782/0.058	0.786/0.054
titanic	0.689/0.033	0.689/0.035	0.69/0.033	0.687/0.033	0.649/0.045	0.689/0.035	0.689/0.035

Tabela B.9: AUC in standardni odklon AUC odločitvenih dreves na domenah z razmerjem razredov 1:100.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.653/0.052	0.613/0.044	0.646/0.059	0.653/0.05	0.634/0.041	0.652/0.061	0.628/0.074
australian-credit	0.809/0.057	0.84/0.034	0.829/0.035	0.848/0.011	0.844/0.026	0.841/0.033	0.851/0.025
blood-transfusion	0.623/0.025	0.64/0.027	0.612/0.009	0.625/0.031	0.612/0.009	0.613/0.027	0.612/0.009
cod-rna	0.707/0.063	0.734/0.075	0.739/0.058	0.742/0.056	0.74/0.06	0.767/0.045	0.74/0.06
contraceptive	0.623/0.047	0.578/0.033	0.592/0.047	0.59/0.043	0.61/0.035	0.571/0.014	0.603/0.041
fourclass	0.894/0.058	0.898/0.062	0.907/0.042	0.915/0.013	0.907/0.042	0.882/0.056	0.907/0.042
gamma-telescope	0.643/0.067	0.663/0.082	0.634/0.06	0.655/0.051	0.637/0.055	0.637/0.07	0.637/0.055
german-credit	0.572/0.035	0.618/0.058	0.578/0.009	0.619/0.049	0.602/0.011	0.624/0.039	0.602/0.045
letter-recognition	0.726/0.04	0.684/0.017	0.725/0.019	0.727/0.057	0.725/0.019	0.718/0.039	0.725/0.019
mammographic-mass	0.824/0.05	0.796/0.021	0.777/0.034	0.811/0.032	0.784/0.029	0.788/0.045	0.781/0.033
musk-v2	0.764/0.064	0.734/0.047	0.748/0.07	0.726/0.041	0.748/0.07	0.772/0.049	0.748/0.07
page-blocks	0.876/0.054	0.869/0.039	0.869/0.036	0.87/0.035	0.869/0.036	0.871/0.036	0.869/0.036
pima-diabetes	0.689/0.066	0.663/0.092	0.682/0.042	0.721/0.045	0.682/0.042	0.686/0.036	0.682/0.042
segmentation	0.912/0.029	0.917/0.021	0.908/0.019	0.923/0.029	0.906/0.02	0.881/0.04	0.906/0.02
spam	0.477/0.039	0.477/0.039	0.586/0.076	0.574/0.069	0.586/0.076	0.585/0.076	0.586/0.076
splice	0.887/0.019	0.906/0.017	0.875/0.014	0.896/0.017	0.869/0.037	0.826/0.023	0.877/0.018
svmguide1	0.934/0.031	0.934/0.031	0.934/0.031	0.93/0.026	0.934/0.031	0.932/0.03	0.934/0.031
svmguide3	0.499/0.001	0.499/0.001	0.5/0.004	0.499/0.001	0.499/0.001	0.499/0.001	0.627/0.049
tic-tac-toe	0.808/0.044	0.8/0.052	0.788/0.045	0.786/0.054	0.808/0.047	0.783/0.068	0.809/0.052
titanic	0.689/0.035	0.689/0.035	0.683/0.032	0.689/0.035	0.687/0.035	0.687/0.035	0.683/0.032

Tabela B.10: Nadaljevanje tabele B.9.

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.597/0.068	0.68/0.075	0.762/0.043	0.683/0.064	0.624/0.055	0.665/0.072	0.645/0.073
australian-credit	0.805/0.122	0.862/0.036	0.837/0.075	0.636/0.169	0.566/0.079	0.798/0.027	0.798/0.027
blood-transfusion	0.636/0.077	0.65/0.068	0.656/0.041	0.631/0.049	0.629/0.039	0.648/0.069	0.648/0.069
cod-rna	0.688/0.041	0.689/0.056	0.752/0.105	0.766/0.068	0.648/0.072	0.652/0.043	0.652/0.043
contraceptive	0.61/0.018	0.61/0.04	0.605/0.052	0.625/0.036	0.572/0.063	0.612/0.05	0.612/0.05
fourclass	0.955/0.03	0.93/0.035	0.875/0.096	0.929/0.038	0.803/0.069	0.952/0.014	0.952/0.014
gamma-telescope	0.64/0.041	0.65/0.07	0.732/0.078	0.564/0.103	0.533/0.052	0.641/0.091	0.641/0.091
german-credit	0.602/0.042	0.632/0.045	0.616/0.059	0.604/0.037	0.637/0.077	0.619/0.007	0.619/0.007
letter-recognition	0.757/0.032	0.528/0.06	0.515/0.049	0.512/0.038	0.502/0.008	0.702/0.036	0.702/0.036
mammographic-mass	0.792/0.056	0.868/0.022	0.843/0.025	0.717/0.188	0.707/0.18	0.829/0.016	0.817/0.029
musk-v2	0.744/0.04	0.785/0.065	0.759/0.045	0.622/0.132	0.532/0.048	0.748/0.07	0.748/0.07
page-blocks	0.823/0.03	0.846/0.06	0.811/0.061	0.63/0.122	0.635/0.128	0.815/0.058	0.815/0.058
pima-diabetes	0.672/0.013	0.702/0.036	0.638/0.119	0.617/0.102	0.588/0.078	0.708/0.054	0.626/0.065
segmentation	0.898/0.035	0.913/0.031	0.875/0.043	0.876/0.091	0.627/0.044	0.914/0.021	0.914/0.021
splice	0.874/0.059	0.874/0.045	0.811/0.094	0.828/0.106	0.709/0.05	0.922/0.032	0.922/0.032
svmguide1	0.876/0.041	0.933/0.022	0.932/0.024	0.801/0.198	0.639/0.202	0.936/0.028	0.936/0.028
svmguide3	0.588/0.025	0.5/0	0.589/0.026	0.501/0.003	0.501/0.003	0.5/0	0.5/0
tic-tac-toe	0.811/0.043	0.82/0.035	0.705/0.038	0.833/0.016	0.657/0.068	0.766/0.042	0.766/0.042
titanic	0.715/0.012	0.718/0.032	0.715/0.012	0.708/0.019	0.655/0.082	0.699/0.017	0.699/0.017

Tabela B.11: AUC in standardni odklon AUC odločitvenih dreves na domenah z razmerjem razredov 1:500.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.593/0.054	0.598/0.093	0.573/0.101	0.657/0.069	0.563/0.067	0.592/0.051	0.585/0.042
australian-credit	0.841/0.05	0.809/0.053	0.847/0.043	0.798/0.027	0.784/0.033	0.809/0.038	0.825/0.055
blood-transfusion	0.594/0.026	0.649/0.063	0.587/0.029	0.648/0.069	0.587/0.029	0.597/0.034	0.587/0.029
cod-rna	0.631/0.055	0.633/0.041	0.666/0.019	0.648/0.045	0.668/0.019	0.689/0.037	0.668/0.019
contraceptive	0.598/0.059	0.602/0.048	0.575/0.015	0.612/0.05	0.532/0.055	0.591/0.033	0.578/0.036
fourclass	0.932/0.065	0.921/0.061	0.948/0.023	0.952/0.014	0.948/0.023	0.929/0.03	0.948/0.023
gamma-telescope	0.642/0.061	0.614/0.073	0.637/0.029	0.641/0.091	0.636/0.029	0.63/0.047	0.637/0.029
german-credit	0.632/0.019	0.619/0.015	0.622/0.017	0.619/0.007	0.622/0.015	0.539/0.048	0.595/0.02
letter-recognition	0.696/0.024	0.682/0.024	0.696/0.037	0.702/0.036	0.696/0.037	0.696/0.043	0.696/0.037
mammographic-mass	0.79/0.067	0.825/0.035	0.777/0.052	0.823/0.031	0.779/0.062	0.717/0.061	0.724/0.027
musk-v2	0.764/0.051	0.744/0.055	0.756/0.052	0.748/0.07	0.756/0.052	0.736/0.05	0.756/0.052
page-blocks	0.787/0.101	0.825/0.054	0.832/0.061	0.815/0.058	0.832/0.061	0.785/0.053	0.832/0.061
pima-diabetes	0.703/0.051	0.645/0.078	0.676/0.03	0.693/0.054	0.676/0.03	0.676/0.028	0.676/0.03
segmentation	0.892/0.017	0.894/0.03	0.897/0.019	0.914/0.021	0.897/0.019	0.914/0.041	0.897/0.019
splice	0.858/0.101	0.935/0.024	0.899/0.053	0.922/0.032	0.901/0.055	0.828/0.045	0.908/0.049
svmguide1	0.941/0.023	0.94/0.022	0.936/0.017	0.936/0.028	0.936/0.017	0.931/0.02	0.936/0.017
svmguide3	0.5/0	0.5/0	0.501/0.003	0.501/0.003	0.5/0	0.5/0.001	0.615/0.045
tic-tac-toe	0.785/0.046	0.77/0.042	0.809/0.059	0.766/0.042	0.819/0.073	0.787/0.044	0.791/0.042
titanic	0.699/0.017	0.699/0.017	0.721/0.031	0.699/0.017	0.72/0.03	0.702/0.019	0.721/0.031

Tabela B.12: Nadaljevanje tabele B.11.

Domena	ReliefFexpRank	InfGain	MDL	Gini	Accuracy	DKM	UniformDKM
adult	0.549/0.045	0.679/0.034	0.706/0.081	0.623/0.049	0.593/0.076	0.649/0.04	0.643/0.045
australian-credit	0.839/0.03	0.87/0.03	0.853/0.091	0.576/0.137	0.53/0.042	0.821/0.032	0.821/0.032
blood-transfusion	0.589/0.03	0.654/0.02	0.659/0.022	0.631/0.029	0.67/0.028	0.578/0.042	0.578/0.042
cod-rna	0.655/0.04	0.673/0.102	0.739/0.103	0.715/0.08	0.627/0.087	0.665/0.052	0.665/0.052
contraceptive	0.585/0.055	0.665/0.078	0.642/0.07	0.664/0.054	0.602/0.061	0.613/0.047	0.613/0.047
fourclass	0.942/0.033	0.954/0.016	0.944/0.02	0.953/0.025	0.87/0.08	0.964/0.024	0.964/0.024
gamma-telescope	0.609/0.026	0.674/0.041	0.663/0.109	0.586/0.086	0.547/0.06	0.635/0.034	0.636/0.036
german-credit	0.609/0.053	0.629/0.045	0.572/0.08	0.622/0.034	0.564/0.071	0.607/0.057	0.607/0.057
letter-recognition	0.727/0.015	0.513/0.041	0.511/0.036	0.513/0.042	0.508/0.026	0.721/0.072	0.721/0.072
mammographic-mass	0.843/0.061	0.861/0.016	0.775/0.028	0.87/0.031	0.818/0.025	0.861/0.039	0.855/0.032
page-blocks	0.789/0.045	0.819/0.1	0.807/0.062	0.582/0.113	0.561/0.059	0.796/0.086	0.796/0.086
pima-diabetes	0.687/0.045	0.725/0.057	0.72/0.065	0.716/0.055	0.681/0.077	0.765/0.034	0.753/0.045
segmentation	0.926/0.032	0.914/0.043	0.915/0.04	0.886/0.1	0.62/0.042	0.921/0.041	0.921/0.041
splice	0.842/0.035	0.795/0.174	0.886/0.036	0.63/0.058	0.614/0.024	0.888/0.033	0.888/0.033
svmguide1	0.887/0.031	0.942/0.042	0.937/0.039	0.882/0.145	0.632/0.126	0.919/0.016	0.916/0.012
svmguide3	0.587/0.033	0.5/0.001	0.602/0.075	0.5/0.004	0.5/0.004	0.5/0.001	0.5/0
tic-tac-toe	0.868/0.041	0.817/0.046	0.734/0.013	0.827/0.019	0.655/0.041	0.793/0.091	0.794/0.09
titanic	0.682/0.013	0.669/0.026	0.682/0.016	0.668/0.023	0.669/0.023	0.684/0.034	0.684/0.034

Tabela B.13: AUC in standardni odklon AUC odločitvenih dreves na domenah z razmerjem razredov 1:1000.

Domena	UniformGini	UniformInf	UniformAccuracy	DistHellinger	DistAUC	DistAngle	DistEuclid
adult	0.559/0.035	0.582/0.053	0.556/0.04	0.646/0.043	0.59/0.042	0.563/0.051	0.567/0.05
australian-credit	0.774/0.076	0.792/0.085	0.813/0.037	0.821/0.032	0.83/0.035	0.829/0.05	0.825/0.081
blood-transfusion	0.581/0.054	0.593/0.041	0.6/0.016	0.578/0.042	0.6/0.016	0.58/0.05	0.6/0.016
cod-rna	0.643/0.081	0.618/0.083	0.625/0.033	0.665/0.052	0.626/0.035	0.633/0.035	0.625/0.033
contraceptive	0.595/0.045	0.622/0.048	0.576/0.009	0.613/0.047	0.567/0.015	0.56/0.045	0.579/0.017
fourclass	0.949/0.071	0.943/0.071	0.954/0.026	0.964/0.024	0.954/0.026	0.931/0.036	0.954/0.026
gamma-telescope	0.618/0.028	0.621/0.035	0.611/0.037	0.636/0.036	0.612/0.038	0.612/0.034	0.612/0.038
german-credit	0.588/0.04	0.562/0.055	0.573/0.032	0.607/0.057	0.577/0.074	0.583/0.028	0.571/0.026
letter-recognition	0.672/0.05	0.686/0.072	0.658/0.017	0.721/0.072	0.658/0.017	0.67/0.008	0.658/0.017
mammographic-mass	0.847/0.037	0.845/0.047	0.803/0.029	0.855/0.032	0.835/0.031	0.817/0.038	0.771/0.021
page-blocks	0.81/0.069	0.809/0.079	0.801/0.077	0.796/0.086	0.801/0.077	0.796/0.068	0.801/0.077
pima-diabetes	0.747/0.046	0.757/0.039	0.728/0.027	0.755/0.036	0.728/0.027	0.734/0.031	0.728/0.027
segmentation	0.905/0.037	0.924/0.031	0.901/0.027	0.921/0.041	0.901/0.027	0.893/0.02	0.901/0.027
splice	0.873/0.018	0.882/0.046	0.862/0.017	0.887/0.033	0.908/0.024	0.827/0.022	0.896/0.018
svmguide1	0.926/0.018	0.933/0.03	0.926/0.018	0.916/0.012	0.926/0.018	0.919/0.016	0.926/0.018
svmguide3	0.5/0.001	0.5/0.001	0.5/0.001	0.5/0.001	0.5/0.001	0.5/0.001	0.621/0.044
tic-tac-toe	0.824/0.056	0.811/0.067	0.82/0.068	0.793/0.091	0.823/0.058	0.829/0.05	0.827/0.053
titanic	0.684/0.034	0.684/0.034	0.684/0.034	0.684/0.034	0.684/0.034	0.695/0.02	0.684/0.034

Tabela B.14: Nadaljevanje tabele B.13.

Literatura

[1] Chih-Chung Chang in Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, št. 2, str. 1–27, 2011. Dostopno na: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] Tomaž Curk, Janez Demšar, Qikai Xu, Gregor Leban, Uroš Petrovič, Ivan Bratko, Gad Shaulsky in Blaž Zupan. Microarray data mining with visual programming. *Bioinformatics*, št. 21, str. 396–398, feb. 2005. Dostopno na: `http://bioinformatics.oxfordjournals.org/content/21/3/396.full.pdf`.

[3] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, št. 7, str. 1–30, 2006.

[4] Thomas G. Dietterich, Michael Kerns in Yishay Mansour. Applying the weak learning framework to understand and improve C4.5. Lorenza Saitta, urednica, v zborniku *Machine Learning: Proceedings of the Thirteenth International Conference (ICML'96)*, str. 96–103. Morgan Kaufmann, San Francisco, 1996.

[5] A. Frank in A. Asuncion. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. Dostopno na: `http://archive.ics.uci.edu/ml`, 2011.

[6] Igor Kononenko in Marko Robnik Šikonja. *Inteligentni sistemi*. UL Fakulteta za računalnistvo in informatiko, Ljubljana, 2010.

[7] University of Toronto. Delve datasets. Dostopno na: `http://www.cs.toronto.edu/~delve/data/datasets.html`, 2011.

[8] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. Dostopno na: `http://www.R-project.org/`.

[9] Marko Robnik Šikonja in Igor Kononenko. Attribute dependencies, understandability and split selection in tree based models. Ivan Bratko in Sašo Džeroski, urednika, v zborniku *Machine Learning: Proceedings of the Sixteenth International Conference (ICML'99)*, str. 344–353. Morgan Kaufmann, 1999.

[10] Marko Robnik Šikonja in Petr Savicky. *CORElearn: CORElearn - classification, regression, feature evaluation and ordinal evaluation*, 2011. R package version 0.9.34. Dostopno na: `http://CRAN.R-project.org/package=CORElearn`.

[11] Peter Savicky in Marko Robnik Šikonja. Evaluation of attributes in imbalanced data sets. Tehnično poročilo, Institute of Computer Science, Academy of Science of Czech Republic, Czech Republic in University of Ljubljana, Faculty of Computer and Information Science, Slovenia, 2011.