

Univerza v Ljubljani  
*Medicinska* fakulteta



Vilma Sem

## Uvrščanje neuravnoteženih visoko razsežnih podatkov v več razredov

Doktorsko delo

Imenovanje mentorja na seji senata dne 28. 10. 2014

Tema doktorske disertacije potrjena na seji senata dne 28. 10. 2014

Komisija za oceno in zagovor imenovana na seji senata dne 21. 11. 2016

Datum zagovora: 5. 1. 2017

Mentorica: izr. prof. dr. Lara Lusa

Jezikovni pregled: Mojca Garntini, prof. slov. j. in knj.

Predsednik komisije: prof. dr. Janez Stare

Član: doc. dr. Rok Blagus

Članica: prof. dr. Katarina Košmelj



# Zahvala

Zahvaljujem se mentorici izr. prof dr. Lari Lusa za usmerjanje pri raziskovalnem delu ter za čas in potrpljenje, ki ga je pri tem potrebovala. Zahvaljujem se dr. Jani Kolar za večletno sodelovanje, iz katerega se je razvila osnovna ideja tega doktorskega dela; zahvaljujem se ji za možnost uporabe spektroskopskih podatkov ter za mnoge strokovne nasvete in spodbude pri mojem delu. Sodelavki doc. dr. Tadeji Kraner Šumenjak se zahvaljujem, da mi je na delovnem mestu omogočila čas, ki sem ga lahko posvetila raziskovalnemu delu v sklopu doktorskega študija ter za spodbude, ko mi je zmanjkovalo motivacije. Mojemu možu se zahvaljujem za razumevanje v času intenzivnega dela in za podporo, ki je ob drugih drobnih dejanjih vključevala tudi dostop do njegovega zmogljivega računalnika, poslušanje predstavitev, kuhanje čaja in sproščujoče pripombe.



## Izvleček

Z različnimi kombinacijami metod predobdelave, zmanjšanja dimenzije podatkov, zmanjšanja vpliva neravnotežja in metod uvrščanja smo zgradili 300 modelov za uvrščanje 45 vrst polimerov na osnovi podatkov bližnje infrardeče spektroskopije. Modele smo ovrednotili z različnimi merami in pristopi za vrednotenje uvrščanja ter izbrali najboljši model, ki napoveduje neznane enote z dovolj visoko stopnjo natančnosti, da je praktično uporaben za identificiranje polimernih materialov v muzejskih zbirkah. Tekom tega postopka smo zbirali podatke o delovanju 16 mer za vrednotenje uvrščanja, za katere je analiza konkordance pokazala visoko stopnjo skladnosti z vrednostjo Kendallovega koeficienta konkordance  $W = 0,95$ .

Kompleksnost pri podatkih bližnje infrardeče spektroskopije, ki se pogosto uporabljajo za določanje kvantitativnih in kvalitativnih lastnosti materialov na različnih področjih, predstavljata visoka razsežnost in močna koreliranost spremenljivk. Pogosti lastnosti, ki problem uvrščanja dodatno zapleteta in sta bili prisotni tudi v našem primeru, sta neuravnoteženost podatkov in veliko razredov. Z namenom oblikovanja praktičnih napotkov za uvrščanje podatkov bližnje infrardeče spektroskopije smo vpliv naštetih lastnosti na uvrščanje raziskali tako na primeru uvrščanja 45 vrst polimerov kot tudi v obsežni simulacijski študiji. Pri tem smo za gradnjo modelov uvrščanja uporabili metode, ki se pogosto uporabljajo pri obdelavi podatkov bližnje infrardeče spektroskopije. Ugotovili smo, da je uporaba predobdelave nujna, saj izboljša rezultate uvrščanja, medtem ko metode za zmanjšanje dimenzije podatkov rezultate večinoma poslabšajo. Za zmanjševanje pristranskosti, ki je posledica neuravnoteženih podatkov, sta se prilagoditev praga za uvrščanje in metoda vsak proti vsakem izkazali za obetajoči, medtem ko metoda večkratnega zmanjšanja večjih razredov, še posebej pri visoki stopnji neravnotežja, ni bila uspešna. Od obravnavanih metod uvrščanja je kljub kršeni predpostavki o neodvisnosti spremenljivk linearna diskriminantna analiza izkazala najboljše rezultate. Najbolj občutljiva na neravnotežje je metoda klasifikacijskih dreves, zaradi česar so rezultati pri tej metodi najslabši.

Za namen simulacij smo predlagali dva nova pristopa generiranja podatkov, ki se po lastnostih dobro približajo realnim podatkom. Ugotovili smo, da sta predlagana pristopa primernejša za raziskovanje delovanja statističnih metod na podatkih bližnje infrardeče spektroskopije kot preprostejši pristopi generiranja podatkov, ki se sicer pogosto uporabljajo v simulacijskih študijah. Pokazali smo, da lahko z napačnimi metodami generiranja podatkov naredimo zavarajoče zaključke.

## Abstract

With the aim to classify 45 kinds of polymer materials 300 classifiers were build by combining different methods of pre-treatment, dimensionality reduction, imbalance bias correction and classification methods. Classifiers were validated by different performance measures and model validation techniques. Overall the best model was chosen, which is capable to predict unknown units with accuracy high enough to be applied for identification of polymer materials of historical collections. Meanwhile data of 16 performance measures were collected and concordance analysis showed high degree of agreement between the performance measures with Kendall coefficient of concordance value as high as  $W = 0.95$ .

High-dimensionality and correlation between variables are the main issues of near-infrared

spectroscopic data, which are often used in qualitative and quantitative analyses of materials from different research fields. Class-imbalance and numerous classes are the properties of data that additionally increase the complexity of classification problem and they are present also in our case. The influence of these properties were investigated in the real case study of 45 polymer materials with a comprehensive simulation study with the purpose to provide practical guidance to the classification of near-infrared spectroscopic data. For this purpose methods frequently used for near-infrared spectroscopic data analysis were used in model learning process in this study. Improved classification performance was achieved when a pre-treatment method was used, while the use of dimensionality reduction methods mostly diminished it. Adjustment of classification threshold and one-against-one method proved promising in attempt to reduce the imbalance bias, whereas multiple down-sizing method was not successful. Among employed classification methods the best results were achieved by linear discriminant analysis even though the assumption of uncorrelated features was unrealistic for near-infrared data. In contrast classification trees had the highest imbalance-sensitivity and consequently the worst performance.

For simulation purposes two new data generation approaches were introduced that resemble real data properties. The proposed approaches were shown to be more appropriate for examination of statistical methods using near-infrared spectroscopic data than simpler data generation approaches, which are often used in simulation studies. Our results show that improper data generation approach may lead to erroneous conclusion.

# Kazalo

<b>1</b>	<b>Uvod</b>	<b>1</b>
1.1	Predstavitev problema . . . . .	1
1.2	Prispevek k znanosti . . . . .	4
1.3	Zgradba doktorskega dela . . . . .	4
<b>2</b>	<b>Pregled literature in uporabljenih pristopov pri uvrščanju NIRS podat-</b>	<b>5</b>
	<b>kov</b>	
2.1	Opis NIRS podatkov . . . . .	5
2.1.1	Absorpcija svetlobe v molekulah . . . . .	5
2.1.2	Spektroskopske meritve . . . . .	6
2.1.3	Absorpcijski pasovi funkcionalnih skupin v NIR območju . . . . .	6
2.1.4	Kemijska zgradba polietilena, polipropilena in polistirena . . . . .	7
2.2	Namen uvrščanja NIRS podatkov . . . . .	9
2.2.1	Predobdelava NIRS podatkov . . . . .	10
2.2.2	Zmanjšanje dimenzije v NIRS podatkih . . . . .	11
2.2.3	Metode uvrščanja pri NIRS podatkih . . . . .	11
2.2.4	Vrednotenje uvrščanja v NIRS podatkih . . . . .	12
2.2.5	Pregled literature o merah za vrednotenje uvrščanja . . . . .	12
2.3	Generiranje NIRS podatkov v dosedanjih študijah . . . . .	16
<b>3</b>	<b>Metode</b>	<b>17</b>
3.1	Dejanski podatki plastik . . . . .	17
3.2	Metode predobdelave in zmanjšanja dimenzije podatkov . . . . .	20
3.2.1	Standardna normalna vektorska transformacija . . . . .	20
3.2.2	Kvantilna normalizacija . . . . .	20
3.2.3	Prvi odvod z metodo Savitzky-Golay . . . . .	21
3.2.4	Izbor spremenljivk z največjo varianco . . . . .	21
3.2.5	Izbor spremenljivk z največjo $F$ -statistiko . . . . .	21
3.2.6	Metoda glavnih komponent (PCA) . . . . .	22
3.3	Metode uvrščanja in vrednotenja modelov uvrščanja . . . . .	22
3.3.1	Metode uvrščanja . . . . .	22
3.3.1.1	Linearna diskriminantna analiza . . . . .	22
3.3.1.2	Metode najbližjega sosedu . . . . .	23
3.3.1.3	Odločitvena drevesa . . . . .	23
3.3.1.4	Metoda podpornih vektorjev . . . . .	24
3.3.2	Pristopi za zmanjšanje neravnotežja pri uvrščanju v več razredov . . . . .	25
3.3.2.1	Pristopa vsak proti vsakemu in vsak proti vsem . . . . .	25

3.3.2.2	Večkratno zmanjšanje večjega razreda . . . . .	25
3.3.2.3	Prilagoditev praga za uvrščanje . . . . .	25
3.3.3	Mere za vrednotenje uvrščanja . . . . .	26
3.3.3.1	Mere vrednotenja uvrščanja za posamezen razred . . . . .	26
3.3.3.2	Mere za vrednotenje uvrščanja v več razredov, izračunane na podlagi kontingenčne tabele delovanja klasifikatorja . . . . .	27
3.3.4	Pristopi za vrednotenje uvrščanja . . . . .	32
3.3.4.1	Večkrat ponovljena stratificirana razdelitev na učno in testno množico (angl. stratified split-sampling) . . . . .	32
3.3.4.2	Prečno preverjanje s pregibanjem . . . . .	32
3.3.5	Metode ocenjevanja skladnosti mer za vrednotenje uvrščanja . . . . .	33
3.3.5.1	Kendallov korelacijski koeficient . . . . .	33
3.3.5.2	Spearmanov korelacijski koeficient . . . . .	33
3.3.5.3	Kendallov koeficient konkordance . . . . .	34
3.3.5.4	Prikaz konkordance z rangi na vzporednih oseh . . . . .	34
3.3.5.5	Konkordančni mehurčni diagram . . . . .	34
3.4	Opis generiranja podatkov . . . . .	35
3.4.1	Generiranje neodvisnih podatkov (IND) . . . . .	35
3.4.2	Generiranje koreliranih spremenljivk z bločno kovariančno matriko (MVNblock) . . . . .	36
3.4.3	Generiranje podatkov iz multivariatne normalne porazdelitve s parametri, ocenjenimi iz realnih podatkov (MVNorig) . . . . .	37
3.4.4	Generiranje podatkov na podlagi teoretičnih absorpcij (ABS) . . . . .	37
3.4.4.1	Trend $t_i(x)$ . . . . .	37
3.4.4.2	Konstanta $C_i$ . . . . .	37
3.4.4.3	Teoretična absorpcija polimera $TA^G(x)$ . . . . .	38
3.4.4.4	Dodatna absorpcijska mesta $RA_i(x)$ . . . . .	39
3.4.4.5	Napaka posamezne meritve $\epsilon(x)$ . . . . .	39
3.5	Nastavitve simulacij . . . . .	39
3.5.1	Uvrščanje v dva in tri razrede . . . . .	39
3.5.1.1	Dva razreda . . . . .	40
3.5.1.2	Trije razredi . . . . .	41
3.5.2	Uvrščanje v 45 razredov . . . . .	41
<b>4</b>	<b>Rezultati</b> . . . . .	<b>43</b>
4.1	Uvrščanje dejanskih podatkov polimerov . . . . .	43
4.1.1	Predstavitev rezultatov modelov uvrščanja vrednotenih s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi . . . . .	44
4.1.2	Predstavitev rezultatov modelov uvrščanja vrednotenih s 500 krat ponovljeno stratificirano razdelitvijo na učno in testno množico . . . . .	51
4.1.3	Predstavitev končnega modela . . . . .	57
4.1.4	Uvrščanje ob uporabi metod za zmanjšanje vpliva neravnotežja . . . . .	61
4.1.4.1	Uvrščanje ob uporabi večkratnega zmanjšanja večjih razredov (MDS) . . . . .	61
4.1.4.2	Uvrščanje ob uporabi metode OAO . . . . .	66
4.1.5	Skladnost mer za vrednotenje uvrščanja . . . . .	68
4.2	Generirani spektri . . . . .	73
4.3	Rezultati simulacij . . . . .	79



4.3.1	SVM z linearno in radialno jedrno funkcijo (Sliki 4.3.1 in 4.3.2) . . .	80
4.3.2	Uvrščanje v dva razreda . . . . .	82
4.3.2.1	Brez predobdelave in brez zmanjšanja dimenzije podatkov (Sliki 4.3.3 in 4.3.4) . . . . .	82
4.3.2.2	Predobdelave spektrov (Slika 4.3.5 in Slika 4.3.6) . . . . .	85
4.3.2.3	Zmanjšanje dimenzije podatkov (Slike 4.3.7, 4.3.8 in 4.3.9) . . . . .	88
4.3.2.4	Večkratno zmanjšanje večjega razreda (Slika 4.3.10) . . . . .	92
4.3.2.5	Povzetek rezultatov pri simulacijah uvrščanja v dva razreda . . . . .	94
4.3.3	Uvrščanje v tri razrede . . . . .	97
4.3.3.1	Brez predobdelave in brez zmanjšanja dimenzije podatkov (Slike 4.3.11, 4.3.12, 4.3.13, 4.3.14 in 4.3.15) . . . . .	97
4.3.3.2	Predobdelave spektrov, zmanjšanje dimenzije podatkov in večkratno zmanjšanje večjega razreda . . . . .	105
4.3.3.3	Izbrane spremenljivke (Slike 4.3.16, 4.3.17, 4.3.18) . . . . .	107
4.3.4	Uvrščanje v 45 razredov . . . . .	111
<b>5</b>	<b>Razprava</b>	<b>113</b>
5.1	Izbrani model za uvrščanje 45 vrst polimerov . . . . .	113
5.2	Generiranje umetnih NIRS podatkov . . . . .	114
5.3	Predobdelave . . . . .	115
5.4	Zmanjšanje dimenzije podatkov in izbrane spremenljivke . . . . .	116
5.5	Metode za zmanjšanje vpliva neravnotežja . . . . .	118
5.6	Metode uvrščanja . . . . .	119
5.7	Skladnost mer za vrednotenje uvrščanja . . . . .	120
<b>A</b>	<b>Slike spektrov polimerov</b>	<b>123</b>
<b>B</b>	<b>Rezultati uvrščanja plastik z metodo najbližjih sosedov z več kot enim sosedom</b>	<b>127</b>
<b>C</b>	<b>Primerjava mer vrednotenja uvrščanja za modele, ovrednotene s pristopom razdelitev na učno množico, ali z uporabo metode MDS</b>	<b>133</b>
<b>D</b>	<b>Rezultati simulacij uvrščanja v tri razrede</b>	<b>137</b>
D.1	Predobdelave spektrov (Sliki D.1, D.2) . . . . .	137
D.2	Zmanjšanje dimenzije podatkov (Slike D.3, D.4, D.5) . . . . .	140
D.3	Večkratno zmanjšanje večjega razreda (Slike D.6, D.7, D.8, D.9) . . . . .	143
<b>E</b>	<b>Pogosto uporabljene krajšave</b>	<b>147</b>



# Poglavje 1

## Uvod

Uvod v delo je vzorčna množica različnih sorazmerno starih plastičnih predmetov in želja, da bi na podlagi meritev bližnje infrardeče spektroskopije znali za stare plastične predmete iz muzejskih zbirk določiti vrsto plastike (oz. vrsto polimera), iz katere so sestavljeni. Na podlagi vzorčnih predmetov smo morali zgraditi model uvrščanja. Zato smo zanje pridobili tako meritve bližnje infrardeče spektroskopije, kot tudi zanesljive podatke o vrsti polimera, iz katerega je bil vsak posamezen predmet sestavljen. Izkazalo se je, da je vrst polimerov, ki so sestavljali vzorčne predmete, zelo veliko, in da so predmeti neenakomerno porazdeljeni med vrste polimerov, kar je v navidez preprost primer uvrščanja vneslo nekaj kompleksnosti. Izgradnjo modela uvrščanja so dodatno zapletle lastnosti pojasnjevalnih spremenljivk (spektroskopske meritve bližnjega infrardečega območja), ki jih je bilo veliko in so med seboj močno korelirale. Srečali smo se torej z uvrščanjem neuravnoteženih visoko razsežnih in močno koreliranih podatkov v več razredov. Z namenom, da bi pripravili tudi praktične napotke za uvrščanje podatkov z opisanimi lastnostmi, smo želeli delovanje metod uvrščanja raziskati s simulacijami. Za namen simulacij pa smo potrebovali umetno generirane podatke. Ker se obstoječi pristopi generiranja podatkov niso dovolj približali realnemu stanju, smo bili primorani razmišljati o novih načinih generiranja podatkov, ki jih v doktorskem delu predstavljamo.

V uvodnem poglavju so predstavljene lastnosti podatkov bližnje infrardeče spektroskopije, problem uvrščanja neuravnoteženih visoko razsežnih in močno koreliranih podatkov v več razredov ter generiranja takšnih podatkov. Sledita predstavitev prispevka k znanosti in povzetek zgradbe doktorskega dela.

### 1.1 Predstavitev problema

Pri spektroskopski metodi posvetimo s svetlobo posamezne valovne dolžine na merjen predmet. Pri tem se nekaj svetlobe odbije, nekaj absorbira in nekaj prepusti. Pri določeni valovni dolžini nato izmerimo delež odbite svetlobe (refleksijska spektroskopija). Deležem odbite svetlobe pri nekem zaporedju valovnih dolžin pravimo spekter odboja. Pri metodi bližnje infrardeče spektroskopije (NIRS, angl. Near Infrared Spectroscopy) so valovne dolžine, pri katerih merimo, iz bližnjega infrardečega (NIR, angl. near infrared) območja, to je med 800 in 2500 nm. Absorpcija (in posledično odboj) svetlobe je tesno povezana s kemijsko

zgradbo osvetljenega materiala. Posamezne skupine atomov v kemijskih spojinah absorbirajo svetlobo pri točno določenih valovnih dolžinah, ki jim rečemo absorpcijski pasovi. V NIR območju se količina absorbirane svetlobe v nekem absorpcijskem pasu porazdeljuje po obliki, ki je med Lorentzovo in Gaussovo krivuljo [1]. Ob tem so absorpcijski pasovi široki in se med seboj prekrivajo, zato jih običajno ne moremo neposredno pripisati določeni atomski skupini, kar oteži določanje kemijske zgradbe merjenega materiala. Uporaba NIRS se je zato razširila šele z razvojem zmogljivejših računalnikov in ustreznih statističnih metod, s katerimi lahko iz NIRS spektrov izluščimo uporabne informacije.

Danes NIRS metoda predstavlja hitro, cenovno ugodno in nedestruktivno raziskovalno metodo [2], ki zahteva minimalno predpripravo vzorcev, zato jo uspešno uporabljajo v raziskavah hrane in agrikulturi [3–5], medicini in farmaciji [2, 6–8], analizi goriv [9, 10], recikliranju plastik [11, 12] in na mnogih drugih področjih. NIRS metoda se uporablja za določanje kemijskih in mehanskih lastnosti snovi [13] ter za njihovo identifikacijo oz. uvrščanje [3–5, 9, 11, 12, 14].

Cilj uvrščanja je zgraditi model, ki na podlagi merjenih (pojasnjevalnih) spremenljivk neznani enoti določi pripadnost razredu. Model se zgradi na podlagi znanih enot (učna množica), za katere ob vrednostih merjenih spremenljivk poznamo tudi dejansko pripadnost razredu. V primeru dveh razredov govorimo o *uvrščanju v dva razreda*, v primeru več kot dveh razredov pa o *uvrščanju v več razredov*. Če velikosti razredov v učni množici niso enake, so podatki *neuravnoteženi*; če je v podatkih veliko spremenljivk (običajno več kot enot), so *visoko razsežni*.

Prvi cilj našega dela je zgraditi model uvrščanja, s katerim bi na podlagi NIRS meritev plastičnih predmetov iz muzejskih zbirk napovedali vrsto plastike (oz. vrsto polimera), iz katere je predmet sestavljen. Pojasnjevalne spremenljivke so bile v našem primeru NIRS meritve oz. deleži odboja svetlobe pri 191 valovnih dolžinah (191 spremenljivk), kar podatke uvrsti med visoko razsežne. Značilnost NIRS spremenljivk je tudi, da so med seboj močno korelirane. V primeru naših podatkov je tako imela več kot petina elementov korelacijske matrike vrednosti večje od 0.9. Razrede so predstavljali različni polimeri, ki jih je bilo v učni množici 45. Med razredi je bilo močno prisotno neravnotežje, saj je bilo število enot v največjem razredu 59, v najmanjšem pa 3. Vsaka od opisanih lastnosti podatkov (visoka razsežnost, koreliranost, veliko število razredov in neravnotežje) vnese v uvrščanje dodatno kompleksnost.

Težava pri uvrščanju visoko razsežnih podatkov je, da nekaterih metod uvrščanja v primeru, ko je število spremenljivk večje od števila enot, ne moremo uporabiti. Večje število spremenljivk predstavlja tudi daljši čas izgradnje modela uvrščanja, čeprav pogosto informacijo, ki je za uvrščanje pomembna, nosijo le nekatere izmed spremenljivk, ostale pa zmanjšujejo učinkovitost modela. Zmanjšana učinkovitost modelov uvrščanja se pogosto kaže kot prepričanje, kar pomeni, da se model ucnim podatkom prilega zelo dobro, a je v primeru testnih podatkov neučinkovit. Probleme visoko razsežnih podatkov pogosto rešujemo z metodami za izbor spremenljivk (pregled metod v bioinformatiki najdemo v članku [15], pregled metod za podatke NIRS pa v [16]).

Visoka koreliranost pojasnjevalnih spremenljivk povzroči kolinearnost [17], ki tako kot pri visoki razsežnosti vodi do singularne kovariančne matrike, ki je vzrok, da nekaterih metod na takšnih podatkih ni mogoče uporabiti ali pa se izkažejo kot neučinkovite. Ob tem se pogosto zgodi, da so v procesu izbora spremenljivk ali grajenja modelov izbrane “lažno” pomembne spremenljivke, ki z relevantnimi spremenljivkami samo korelirajo, kar vodi do slabe učin-

kovitosti modelov. Ta pojav je opisan in za nekatere metode tudi s simulacijami raziskan v [18, 19]. Težave, povezane z visoko koreliranostjo, rešujemo z ustreznim izborom spremenljivk ali z metodami, ki zmanjšajo dimenzijo podatkov tako, da iz obstoječih spremenljivk izračunajo tako imenovane latentne spremenljivke, ki so med seboj pogosto neodvisne, kot sta npr. metoda glavnih komponent (PCA, angl. Principal Component Analysis) in metoda delnih najmanjših kvadratov (PLS, angl. Partial Least Squares).

Nekatere metode uvrščanja, ki so bile razvite za uvrščanje v dva razreda, lahko posplošimo, da delujejo za uvrščanje v več razredov [20–22], kar običajno močno poveča kompleksnost algoritma. Za nekatere metode takšna posplošitev za uvrščanje v več razredov sploh ne obstaja. Zato so razvili tudi pristope, ki problem uvrščanja v več razredov razdelijo na več problemov uvrščanja v dva razreda: vsak proti vsem [23], vsak proti vsakemu [24] in P proti Q [25] (kritičen pregled teh metod in primerjavo na primeru uvrščanja z metodo podpornih vektorjev najdemo v študijah [23, 26]).

Več študij je pokazalo, da so metode za uvrščanje na neravnotežje občutljive in da so rezultati pristranski v prid večjemu razredu [27–29]. He in Garcia [27] sta predstavila pregled pristopov, ki so jih razvili za zmanjšanje problema neravnotežja. Problem neravnotežja je bil sistematično raziskan za različne klasifikatorje za primer uvrščanje v dva razreda [27–31]. Uvrščanja v več razredov z neuravnoteženimi podatki pa se bežno dotaknejo le nekatere od njih. Zhou in Liu [32] sta na primeru umetnih nevronske mreže empirično pokazala, da nekatere metode za neuravnotežene podatke, ki dobro delujejo na primeru uvrščanja v dva razreda, za uvrščanje v več razredov niso učinkovite. Za umetne nevronske mreže so bili v študiji [25] natančno raziskani tudi pristopi vsak proti vsem, vsak proti vsakemu in P proti Q na primeru empiričnih neuravnoteženih podatkov. Avtorja članka [33] ugotavljata, da je področje uvrščanja neuravnoteženih podatkov v več razredov še slabo raziskano. V študiji sta sistematično obravnavala uvrščanje nizko razsežnih podatkov s petimi različicami algoritma *AdaBoost* pri obeh primerih neuravnoteženosti podatkovne baze z več razredi: *multimajoriti* (en manjši razred in več večjih razredov) in *multiminoriti* (en večji razred in več manjših razredov). Problem visoko razsežnih neuravnoteženih podatkov je obravnavan v študiji [34], kjer so se ukvarjali s satelitskimi hiperspektralnimi podatki zemeljskega površja, ki so jih uvrščali v 16 razredov. Za zmanjšanje vpliva neravnotežja so uporabili naključno zmanjšanje večjih razredov (RUS, angl. Random Undersampling) in metodo SMOTE (SMOTE, angl. Synthetic Minority Over-sampling Technique), za zmanjšanje razsežnosti podatkov pa metodo PCA.

Veliko študij je primerjalo različne metode za uvrščanje podatkov NIRS [10, 35–38]. Največkrat uporabljene metode pri obdelavi NIRS podatkov so bile: diskriminantna analiza delnih najmanjših kvadratov (PLS-DA, angl. Partial Least Squares Discriminant Analysis) [36, 37], linearna diskriminantna analiza [10, 35, 37, 38], kvadratna diskriminantna analiza [10, 38], regularizirana diskriminantna analiza [10], metoda  $k$ -najbližjih sosedov [10, 37, 38], metoda podpornih vektorjev [10, 37], mehko neodvisno modeliranje podobnosti po razredih (SIMCA, angl. Soft Independent Modeling of Class Analogy) [10, 35, 39–41], različne oblike umetnih nevronske mreže (verjetnostna nevronska mreža – PNN, angl. Probabilistic Neural Network in nevronska mreža z vzvratnim širjenjem napake – BPNN, angl. Back Propagation Neural Network) [10, 37, 41] in odločitvena drevesa (CART, angl. Classification And Regression Trees) [42, 43].

V zadnjem času je bilo na temo uvrščanja neuravnoteženih podatkov objavljenih veliko študij predvsem s področja strojnega učenja [27, 30, 33, 44, 45], vendar so se v večini omejele

na dva razreda ali na nizko razsežne podatke. Naš nadaljnji cilj je bil zato s simulacijami raziskati lastnosti metod uvrščanja v več razredov z namenom, da bi pripravili praktične napotke za uvrščanje NIRS podatkov in bi, če bi bilo mogoče, izboljšali zgrajeni model za uvrščanje plastik. Zaradi specifičnosti obravnavanih podatkov s preprostimi simulacijskimi metodami ni bilo mogoče simulirati realističnih korelacijskih struktur med spremenljivkami. V literaturi smo sicer našli več poskusov simuliranja NIRS podatkov [46–55], ki pa se niso dovolj približali dejanskemu stanju podatkov v naši študiji in zato niso bili primerni za simuliranje uvrščanja velikega števila materialov. Ker obstoječih simulacijskih idej iz literature nismo mogli uporabiti za proučevanje zelenih lastnosti, smo predlagali nove pristope simuliranja NIRS podatkov. Podatke smo simulirali na štiri različne načine in rezultate primerjali z realnimi podatki v študiji uvrščanja neuravnoteženih podatkov v dva, tri in, v omejenem obsegu, tudi 45 razredov.

## 1.2 Prispevek k znanosti

V doktorskem delu je predstavljena rešitev problema uvrščanja zgodovinskih plastik v 45 razredov, ki je po našem vedenju prva za uporabnika sprejemljiva rešitev uvrščanja zgodovinskih plastik na podlagi NIRS. Rešitev je bil model uvrščanja, ki smo ga izbrali med 300 modeli uvrščanja, ki smo jih zgradili z različnimi kombinacijami metod za uvrščanje, predobdelavo podatkov, zmanjšanje dimenzije podatkov in zmanjšanje vpliva neravnotežja. Modele smo ovrednotili s 16 merami za vrednotenje uvrščanja.

V doktorskem delu sta predlagana dva nova pristopa umetnega generiranja NIRS podatkov, ki sta uporabna za raziskovanje lastnosti različnih statističnih metod na podatkih NIRS, ki so kategorizirani v skupine.

Na podlagi simulacij je bila za linearno diskriminantno analizo, odločitvena drevesa in metodo podpornih vektorjev narejena sistematična študija, ki proučuje lastnosti uvrščanja podatkov bližnje infrardeče spektroskopije v dva, tri in, v omejenem obsegu, tudi 45 razredov ob različnih stopnjah neravnotežja in ob uporabi metod za predobdelavo podatkov, zmanjšanje dimenzije podatkov in zmanjšanje vpliva neravnotežja.

## 1.3 Zgradba doktorskega dela

V uvodu je na kratko predstavljen v doktorskem delu obravnavan problem. V drugem poglavju so predstavljene specifične lastnosti NIRS podatkov ter pregled literature o uporabljenih načinih reševanja problema uvrščanja s področja bližnje infrardeče spektroskopije. V tretjem poglavju so opisane v študiji uporabljene metode. V četrtem poglavju pa so predstavljeni rezultati analiz. V zadnjem poglavju so povzete in kritično ovrednotene glavne ugotovitve.

## Poglavje 2

# Pregled literature in uporabljenih pristopov pri uvrščanju NIRS podatkov

V razdelku 2.1 so predstavljeni tisti temeljni pojmi o absorpciji svetlobe v NIR območju, ki so potrebni za razumevanje predlagane metode umetnega generiranja NIRS podatkov v razdelku 3.4.4. V razdelku 2.2 je pregled literature s področja uvrščanja NIRS podatkov in pregled nekaterih objav s področja mer za vrednotenje uvrščanja v več razredov. Pregled literature, kjer so se ukvarjali z umetnim generiranjem NIRS podatkov, najdemo v razdelku 2.3.

### 2.1 Opis NIRS podatkov

#### 2.1.1 Absorpcija svetlobe v molekulah

Svetloba je elektromagnetno valovanje, ki obsega valovne dolžine od velikosti gama žarkov, preko rentgenskih valov, ultravijolične, vidne in infrardeče svetlobe do mikrovalov in najdaljših, radijskih valov. Infrardeče območje podrobneje delimo na bližnje (NIR), srednje in daljne infrardeče območje. Bližnje infrardeče območje zavzema valovne dolžine takoj za vidno svetlobo, to je od 800 do 2500 nm. Svetloba je sestavljena iz delcev (fotonov), ki prenašajo količino energije ( $E$ ), povezano s točno določeno frekvenco (Planckov zakon [56]):

$$E = h \cdot \nu = h \cdot \frac{c}{\lambda},$$

kjer je  $h$  Planckova konstanta,  $\nu$  frekvenca,  $c$  hitrost svetlobe in  $\lambda$  valovna dolžina.

Energijo molekule lahko opišemo kot vsoto energije, povezane s stanji elektronov ter rotacijske in vibracijske energije. Ko fotoni svetlobe, ki jo ustvarja spektrometer, trčijo ob molekulo merjenega materiala, lahko pride do absorpcije svetlobe, pri čemer molekula preide v višje energetske stanje (vzbujeno stanje). Rotacijska energija molekule je posledica vrtenja molekul in frekvence, ki lahko rotacijo vzbujajo, ustrezajo mikrovalovom. Frekvence, ki lahko vzbujajo elektrone, da prehajajo v višja elektronska stanja, ustrezajo ultravijoličnim

valovom in deloma vidni svetlobi. Frekvence infrardečega območja pa vzbujajo vibracijska nihanja, to so raztezanja kemijskih vezi in spremembe kotov med kemijskimi vezmi, pri čemer morajo biti izpolnjeni določeni pogoji, povezani z dipolnim momentom in simetrijo molekule (za razlago teh pojmov priporočamo [57]). Vibracijskemu nihanju katerekoli vezi med atomi ustreza le točno določena frekvenca. Če molekula svetlobo s to frekvenco absorbira, preide na prvo višje energetske stanje oz. v osnovno nihanje. Pri prehodu molekule iz osnovnega v drugo višje energetske stanje nastane prvi nadton, pri prehodu iz osnovnega v tretje višje energetske stanje drugi nadton itd. Če bi bilo nihanje med atomi v molekuli harmonično, bi prvi nadton nastal pri dvakratniku frekvence osnovnega nihanja, drugi nadton pri trikratniku te frekvence itd. Vendar vibracijska nihanja v molekuli niso harmonična, zato je izračun frekvenc nadtonov bolj kompleksen, za dvoatomno molekulo se mu še najbolj približa model Morsejevega nihanja [58]. Ob osnovnih nihanjih in nadtonih lahko ob prodoru svetlobne energije v molekulo nastanejo tudi kombinacije vibracij, ki so lahko prvega ali drugega reda. Kombinacije prvega reda nastanejo, kadar sta hkrati vzbujeni dve osnovni vibracijski nihanji z zelo podobnima frekvencama. Kombinacije drugega reda so znane pod imenom Fermijeve resonanca [59] in so posledica medsebojnega vpliva enega osnovnega vibracijskega nihanja in enega nadtona ali kombinacije prvega reda. Posledice Fermijeve resonance se kažejo kot zamik frekvence in ojačanje pričakovane amplitude vibracije [57]. S številom atomov v molekuli se število različnih vibracij hitro povečuje, s tem se povečuje tudi kompleksnost določevanja frekvenc, tako osnovnih vibracij kot tudi nadtonov in kombinacij vibracij.

Absorpcija osnovnih vibracijskih nihanj vezi je najmočnejša in se za večino molekul nahaja v srednjem infrardečem območju (MIR, angl. mid infrared, 3000–8000 nm). V bližnjem infrardečem območju (NIR, angl. near infrared, 800–2500 nm) je opazna absorpcija višjih harmoničnih tonov (oz. nadtonov) in kombinacij prvega in drugega reda. Višji kot je red nadtona, manjša je moč absorpcije in pri nižjih valovnih dolžinah se pojavi [60, 61]. To pomeni, da je absorpcija osnovne vibracije najmočnejša, po moči ji nato sledi prvi višji harmonični ton, nato drugi itd. Zato je v spektralnih meritvah infrardečega območja opazen trend: z manjšanjem valovnih dolžin se manjša moč absorpcij.

### 2.1.2 Spektroskopske meritve

Ker absorbirane svetlobe ne moremo neposredno izmeriti, s spektrometri merimo amplitudo od vzorca odbite (refleksijska spektroskopija) ali skozi vzorec prepuščene (transmisijska spektroskopija) svetlobe v odvisnosti od valovne dolžine. Rezultat meritev imenujemo refleksijski oz. transmisijski spekter. Odboj in prepustnost sta običajno podana relativno glede na vpadno svetlobo. Na splošno velja, da je količina vpadne svetlobe enaka vsoti v vzorcu absorbirane, iz vzorca odbite in skozi vzorec prepuščene svetlobe. Pod določenimi "idealnimi" pogoji, ki so v praksi redko vsi izpolnjeni, lahko povezavo med absorpcijo, odbojem in prepustnostjo izrazimo z zakoni kvantne fizike, kot sta Beer-Lambertova ali Schuster-Kubelka-Munkova teorija [1].

### 2.1.3 Absorpcijski pasovi funkcionalnih skupin v NIR območju

Funkcionalna skupina je atom ali skupina atomov, ki je za posamezno organsko spojino značilna in določa njene tipične lastnosti. Vsaka funkcionalna skupina absorbira elektromagnetno valovanje v zanje značilnih absorpcijskih pasovih, to je pri točno določenih valovnih



dolžinah. Teoretična oblika absorpcijske krivulje v enem absorpcijskem pasu je med Lorentzovo in Gaussovo [1]. Absorpcijske pasove pri posamezni funkcionalni skupini ter višino in širino absorpcijskih krivulj je zelo težko določiti, saj frekvence v NIR območju ustrezajo predvsem energiji nadtonov in kombinacij C – H, N – H in O – H vezi. Ob tem da so te absorpcije veliko šibkejše od osnovnih vibracij, se absorpcijski pasovi med seboj tudi prekrivajo, kar je vzrok za veliko kompleksnost NIR spektrov.

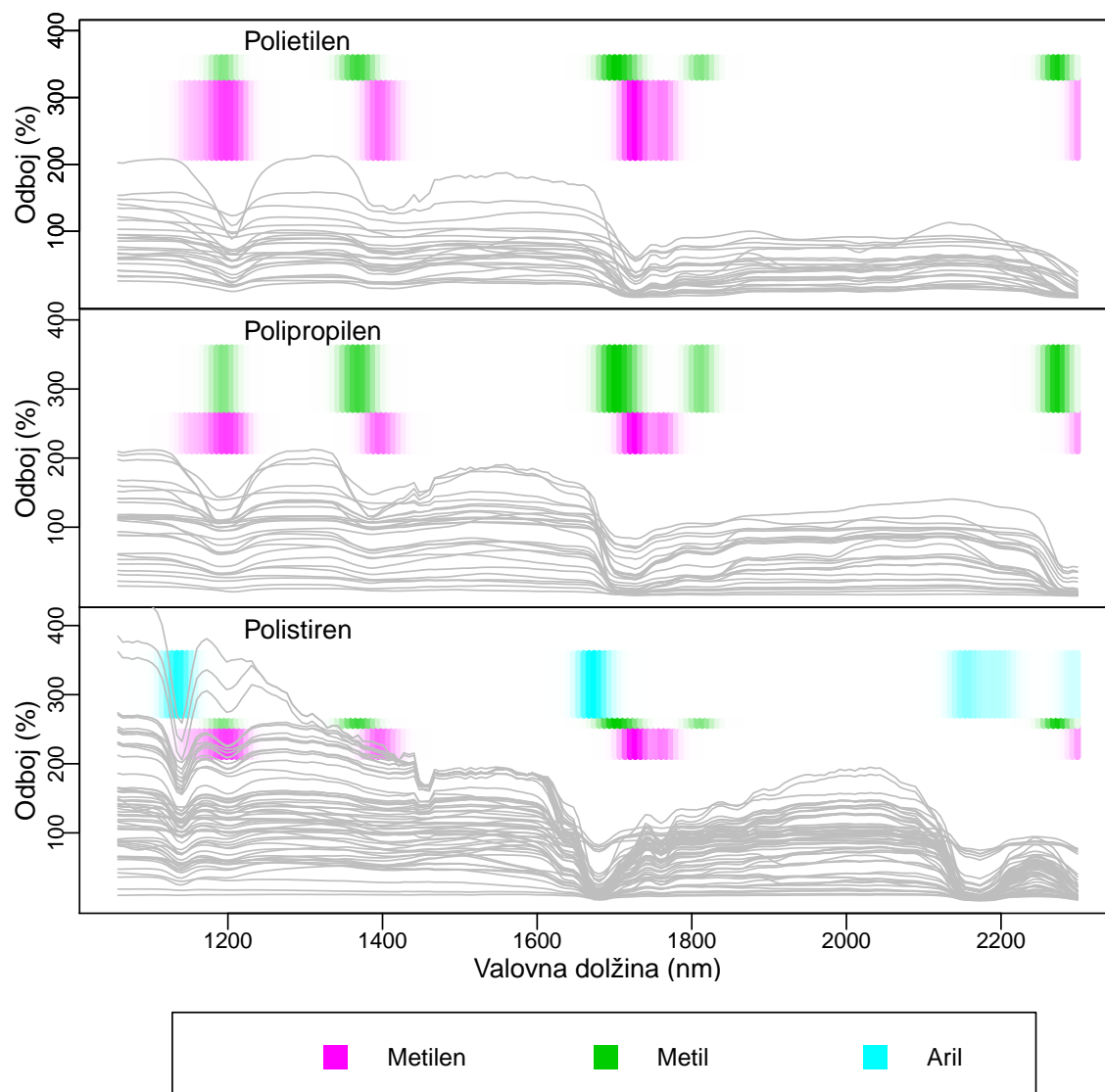
Mesto vrha absorpcijske krivulje v absorpcijskem pasu je sicer mogoče teoretično izračunati po zakonih kvantne fizike, vendar se v spektrih vedno pojavijo odstopanja izmerjenih absorpcij od izračunanih. Na zamik mesta absorpcije vplivajo predvsem vodikove vezi in drugi sestavni deli molekule, v kateri je funkcionalna skupina vezana (angl. neighbouring group effect) [60]. Še težje kot mesto absorpcije je določiti moč absorpcije (višino absorpcijske krivulje), na katero ob vodikovih vezeh in drugih sestavnih delih molekule vpliva še Fermijeva resonanca [59, 60]. Absorpcija na določenem mestu je močnejša, če je funkcionalna skupina, ki absorpcijo povzroča, v molekuli vezana večkrat [57, 61]. Ob vseh omenjenih dejavnikih na absorpcijo funkcionalnih skupin v NIR območju vplivata tudi stopnja kristaliničnosti in razvejanost molekul ter pogoji v prostoru, kot je npr. temperatura. Zaradi zapletenih absorpcij v NIR območju so absorpcijska mesta posameznih funkcionalnih skupin določena empirično, rezultati meritev za iste funkcionalne skupine pa se lahko močno razlikujejo.

V literaturi so ponekod objavljena le široka območja, v katerih lahko pričakujemo absorpcijska mesta določene funkcionalne skupine, kot npr. v [62]. V nekaterih drugih virih so kot mesta absorpcij objavljene posamezne valovne dolžine, ki naj bi predstavljale vrhove absorpcijskih krivulj (npr. [57, 63–65]). Višina absorpcijskih krivulj oz. moč absorpcije je podana zelo redko in nikoli s točnimi vrednostmi.

#### 2.1.4 Kemijska zgradba polietilena, polipropilena in polistirena

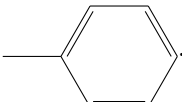
V nadaljevanju, kjer se bomo ukvarjali z umetnim generiranjem spektrov, bomo osredotočeni na spektre treh vrst polimerov: polietilena (PE), polipropilena (PP) in polistirena (PS). Te tri vrste polimerov so v naši zbirki plastičnih predmetov, ki je natančno predstavljena v razdelku 3.1, in tudi sicer v zbirkah plastičnih predmetov v muzejih med pogostejšimi. Ker so spektri v tesni povezavi s kemijsko zgradbo merjenega vzorca, bomo za boljše razumevanje metod generiranja spektrov v tem razdelku predstavili kemijsko zgradbo omenjenih polimerov.

Na Sliki 2.1.1 lahko vidimo dejanske spektralne meritve predmetov iz vzorčne zbirke, opisane v razdelku 3.1, ki so bili narejeni iz materialov PE, PP in PS. Vsaka navidezna krivulja je bila pridobljena na osnovi meritev enega predmeta (ene enote) in je sestavljena iz diskretnih točk, ki predstavljajo delež odbite svetlobe pri posamezni valovni dolžini. Delež odbite svetlobe ponekod presega 100 %, za kar obstaja več razlogov. Spektrometer se na začetku merjenja umeri po tako imenovanem “belem standardu”, za katerega se predpostavi 100 % odboj in pri nadaljnjih meritvah služi kot referenčna meritev. V resnici pa material s 100 % odbojem ne obstaja. Odboj svetlobe na standardu dodatno zmanjša tudi prisotnosti nečistoč. Pri našem delu je bil uporabljen standard *Spectralon*, več o njegovih lastnostih je navedeno v [66]. Drugi razlog za izmerjeni odboj nad 100 % je lahko svetloba iz okolice, ki jo je merilna sonda zajela ob svetlobi, odbiti od merjenega materiala. Ta razlog je še posebej verjeten pri transparentnih objektih.



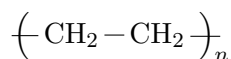
Slika 2.1.1: Merjeni spektri vzorcev polietilena, polipropilena in polistirena iz vzorčne množice podatkov, ki so opisani v razdelku 3.1, in teoretične absorpcije metilne, metilenske in arilne funkcionalne skupine, pridobljene iz literature [57, 63–65, 67]. Močnejša barva predstavlja močnejšo absorpcijo, večji simbol pa predstavlja večji delež absorpcije pripadajoče funkcionalne skupine. Deleži absorpcij po funkcionalnih skupinah so le približni. Merjenih enot PE, PP in PS je bilo  $n_{PE} = 26$ ,  $n_{PP} = 27$  in  $n_{PS} = 59$ .

Polimere polietilen (PE), polipropilen (PP) in polistiren (PS) sestavljajo tri funkcionalne skupine z naslednjo kemijsko zgradbo:

- metilen  $-\text{CH}_2-$ ,
- metil  $-\text{CH}_3$  in
- aril .

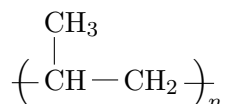
Predvidene absorpcije metilenske, metilne in arilne funkcionalne skupine v NIR območju, pridobljene iz literature [57, 63–65, 67], so prikazane na sliki 2.1.1.

### Polietilen (PE)



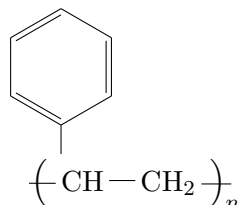
PE je sestavljen iz zaporedno vezanih več 1000 metilenskih skupin  $-\text{CH}_2-$ , veriga pa se zaključí z metilno skupino ( $-\text{CH}_3$ ). Delež metilnih skupin je lahko v primeru razvejane verige (LDPE, angl. Low-Density Polyethylene) sicer nekoliko večji, kljub temu pa v NIR spektru polietilena pričakujemo opaznejšo absorpcijo v območjih metilenske kot v območjih metilne skupine [57, 61].

### Polipropilen (PP)



PP je sestavljen iz približno enakega števila metilne in metilenske skupine. Če so molekule polipropilena razvejane, je vsebnost metilnih skupin večja, ob tem metilna skupina vsebuje tri  $\text{C} - \text{H}$  vezi, medtem ko metilenska skupina vsebuje le dve, zato lahko v NIR spektru pričakujemo nekoliko opaznejšo absorpcijo v območju metilne kot metilenske skupine.

### Polistiren (PS)



PS sestavlja približno enako število arilne in metilenske skupine, na obeh koncih pa se veriga zaključí z metilno skupino, vendar lahko zaradi istega razloga kot pri PP in tudi zato, ker imajo nadtoni  $\text{C} - \text{H}$  vezi iz arilne skupine večjo moč absorpcij kot nadtoni  $\text{C} - \text{H}$  vezi iz metilenske skupine [68], pričakujemo močnejšo absorpcijo v območjih arilne skupine.

## 2.2 Namen uvrščanja NIRS podatkov

Ker so instrumenti, ki so za merjenje NIR spektrov potrebni, prenosljivi, sama metoda pa je hitra in ne poškoduje merjenega vzorca, so NIR metodo poskusili uporabiti na najrazličnejših področjih. Na nekaterih področjih se metoda izkaže za dovolj zanesljivo, da jo lahko uporabljajo namesto dragih in dolgotrajnih analitičnih postopkov, ki običajno zahtevajo predpripravo vzorca, pri kateri se vzorec tudi poškoduje ali celo uniči. Njena uporaba je verjetno najpogostejša v raziskavah s področja varne hrane. Pregled raziskav, kjer so z NIR metodo klasificirali pristnost hrane kot npr. pristnost različnih vrst mesa, pristnost beljakovin v jogurtu, pristnost geografskega porekla vina ipd., najdemo v [69]. S pomočjo uvrščanja NIRS podatkov so uspešno ločili tudi ponarejeno mleko od pravega [70], med glede na botanično poreklo [71] in olivno olje glede na geografsko poreklo [72]. Prav tako

so bili uspešni pri uvrščanju rastlinskih olj glede na rok trajanja [73] in uvrščanju cigaret glede na znamko [14]. Modeli uvrščanja NIRS podatkov so bili uporabljeni tudi za klasifikacijo prsti [74], ločevanje bakterij [75], ločevanje poškodovanih in celih semen [5], določanje sladkorja v sladkornem trsu [76], klasifikacijo goriv [35] ter v medicini za diagnosticiranje raka na prsih [77] ipd.

Kot smo opisali že v uvodu, je uvrščanje NIRS podatkov kompleksen problem zaradi visoko razsežnih podatkov in koreliranosti pojasnjevalnih spremenljivk, pogosto pa je prisotno tudi neravnotežje. Ob tem v našem primeru dodatno kompleksnost prinese veliko število razredov. Uvrščanje visoko razsežnih neuravnoteženih podatkov je bilo že natančno raziskano za uvrščanje v dva razreda [29], kjer so pokazali, da so nekateri klasifikatorji, ki se pogosto uporabljajo v primeru visoko razsežnih podatkov zelo občutljivi na neravnotežje. Ob tem so ugotovili, da se pristranskost v prid večjemu razredu, ki je posledica neravnotežja, v primeru visoko razsežnih podatkov še poveča. Izbor spremenljivk je v večini primerov ugodno vplival na rezultate uvrščanja. V [45] so raziskovali vpliv visoko razsežnih neuravnoteženih podatkov na uvrščanje v dva razreda tudi ob prisotnosti nizko koreliranih spremenljivk. V primeru, ko med razredoma ni bilo razlik v izraženosti spremenljivk, se je problem neravnotežja ob prisotnosti koreliranih spremenljivk zmanjšal, v primeru, ko so bile med razredoma razlike, pa ne.

Z uvrščanjem neuravnoteženih podatkov v več razredov so se ukvarjali v [33], pri tem pa niso obravnavali problema visoke razsežnosti. Ugotovili so, da je problem neravnotežja večji ob prisotnosti večjega števila večjih razredov (angl. *multimajority*), kot ob prisotnosti večjega števila manjših razredov (angl. *multiminority*). Preizkusili so metodo povečanja manjših razredov, ki ni uspešno odpravila problema neravnotežja, saj je povzročila preprileganje manjših razredov. Za metodo zmanjšanja večjih razredov so ugotovili, da je občutljiva na število manjših razredov in da zmanjša napovedno točnost večjih razredov. Najbolj optimistične rezultate so dobili z uporabo algoritma AdaBoost.NC [78]. Za zmanjšanje vpliva neravnotežja pri uvrščanju v več razredov so preizkusili tudi metodo vsak proti vsem in uteženo metodo vsak proti vsem, pri kateri so kot uteži upoštevali deleže razredov v osnovni učni množici. Utežena metoda vsak proti vsem se je sicer izkazala za uspešnejšo, a kljub temu ni izboljšala rezultatov osnovnega modela uvrščanja z vsemi vključenimi razredi naenkrat.

Pri uvrščanju NIRS podatkov se je uveljavila analiza podatkov v naslednjih korakih: (1) predobdelava podatkov, (2) zmanjšanje dimenzije podatkov, (3) izgradnja modela uvrščanja in (4) ovrednotenje modela uvrščanja. V nadaljevanju je pregled literature za najpogostejše uporabljene metode na vsakem od naštetih korakov.

### 2.2.1 Predobdelava NIRS podatkov

Metode predobdelave spektrov uporabljajo za zmanjšanje variabilnosti meritev, ki ni posledica kemijske strukture merjenega predmeta in je v spektroskopiji pogosto poimenovana kot šum (angl. *noise*). Vzroki takšne variabilnosti so povezani z napakami merilne naprave in fizičnimi lastnostmi predmeta, kot so razporejenost delcev v merjeni snovi, površina in barva predmeta ipd. V spektrih se takšen šum pogosto kaže kot zamik, razteg/skrčitev spektrov v navpični smeri ter v obliki manjših odmikov od dejanskih vrednosti pri posamezni valovni dolžini. Metode za predobdelavo NIRS podatkov uporabljajo z namenom, da bi zmanjšali to neželeno variabilnost. Izkazalo se je, da uporaba predobdelave običajno bistveno izboljša

rezultate uvrščanja. V [79] najdemo pregled metod za predobdelavo spektrov. Najpogosteje uporabljena metoda predobdelave je standardna normalna vektorska transformacija (SNV, angl. Standard Normal Variate), ki je bila uporabljena npr. v [5, 37, 70] ali njej podobna multiplikativna korekcija razpršenosti (MSC, angl. Multiplicative Scatter Correction). Včasih je kot predobdelava uporabljeno le glajenje, kot npr. glajenje z metodo Savitsky-Golay v [74], zelo popularna pa je predobdelava s prvim ali drugim odvodom, ki je običajno združena z glajenjem [71, 75]. Enovitega odgovora o tem, katera od metod za predobdelavo spektrov je najbolj učinkovita, ni [79]. Zato raziskovalci pogosto vsaj v začetnem delu raziskave preizkusijo več metod in se na podlagi rezultatov odločijo, katero metodo bodo uporabili v nadaljevanju.

### 2.2.2 Zmanjšanje dimenzije v NIRS podatkih

Cilj zmanjšanja dimenzije podatkov je zmanjšati kompleksnost postopka izgradnje modela uvrščanja ter povečati njegovo učinkovitost in interpretabilnost tako, da v model uvrščanja vključimo le spremenljivke, ki nosijo za uvrščanje relevantno informacijo, oz. iz modela izključimo spremenljivke, ki predstavljajo le dodaten šum. Z naraščanjem števila visoko razsežnih podatkov se je razvila tudi vrsta metod za zmanjšanje dimenzije podatkov (pregled metod v kemometriji najdemo v [80], pregled metod pri NIRS pa v [81, 82]). V članku [83] je izpostavljeno, da določene metode delujejo dobro le za določeno vrsto podatkov, medtem ko lahko pri drugačnem problemu z uporabo istih metod dobimo zavajajoče rezultate in da se raziskovalci v takšni množici različnih metod ne znajdejo. Morda je tudi to razlog, da raziskovalci na področju NIRS za zmanjšanje dimenzije podatkov (in posledično tudi koreliranosti) najpogosteje uporabljajo metodi PCA ali PLS, kjer nove spremenljivke (glavne komponente) izračunamo kot linearno kombinacijo originalnih spremenljivk tako, da vanje zajamemo kar največ informacije iz originalnih spremenljivk. Modele uvrščanja se potem zgradi na nekaj glavnih komponentah. Primere uporabe metod PCA in PLS na NIRS podatkih najdemo v [35, 37, 70, 71, 74, 76]. Metoda PLS je najpogosteje uporabljena v kombinaciji z metodo uvrščanja LDA (PLS-DA) [3, 5, 72, 73, 75]. V nekaterih študijah so dimenzijo podatkov zmanjšali tako, da so s povprečjem združili meritve pri zaporednih valovnih dolžinah [35, 74] ali izbrali spremenljivke (valovne dolžine) glede na vrednost F-statistike [36]. Na nekaterih področjih se je za izbor spremenljivk uveljavil tudi genetski algoritem [84]. Slabost navedenih metod za zmanjšanje dimenzije podatkov je, da je izbrane spremenljivke oz. glavne komponente težko interpretirati, zato so razvili tudi metode, ki v model uvrščanja vključijo podatek o znanih absorpcijskih pasovih [85].

### 2.2.3 Metode uvrščanja pri NIRS podatkih

Rezultati uvrščanja neuravnoteženih podatkov so praviloma pristranski v prid večjemu razredu. Čeprav je problem uvrščanja neuravnoteženih podatkov v literaturi že nekaj časa znan [28], v mnogih študijah s področja uvrščanja NIRS podatkov, kljub prisotnosti neravnotežja, ni bil obravnavan [36, 73–75]. Kjer so se problema neravnotežja zavedali, so ga pogosto reševali v postopku validacije na način, da so v učno množico izbrali za vsak razred enako število enot, v testni množici pa so bile preostale enote kot npr. v [5]. Problema neravnotežja so se nekateri lotili tudi bolj inovativno: v [37] so zaradi neravnotežja izbrali metodo uvrščanja *support vector data description* (SVDD), ki se je razvila iz metode podpornih vektorjev in bolje prepozna tudi manjše razrede, v [86] so razvili model uvrščanja

v dveh stopnjah, kjer so v prvi stopnji izločili delež enot iz večjega razreda in tako zmanjšali neravnotežje na drugi stopnji uvrščanja. Vpliva neravnotežja pri tem niso popolnoma odpravili, ob tem pa se uporabljenega pristopa ne da posplošiti na druge primere.

#### 2.2.4 Vrednotenje uvrščanja v NIRS podatkih

Vrednotenje modelov uvrščanja je izrednega pomena, saj lahko z nepravilnim vrednotenjem dobimo zavajajočo informacijo o zgrajenem modelu. Mere in pristopi za vrednotenje uvrščanja so natančno predstavljeni v razdelkih 3.3.3 in 3.3.4. V študijah, ki vključujejo podatke NIRS, je kot pristop za vrednotenje uvrščanja pogosto uporabljena enkratna razdelitev množice podatkov na učno in testno množico, pri čemer je model uvrščanja zgrajen na učni množici in ovrednoten na testni množici, kot npr. v [3, 5, 71, 74, 75]. Ta pristop vrednotenja uvrščanja je problematičen v primeru majhnega števila podatkov, saj je ob razdelitvi na učno in testno množico učna množica zelo majhna, kar zmanjšuje učinkovitost uvrščanja. Ob tem je ocena uvrščanja pridobljena na tak način močno odvisna od načina razdelitve množice in ima zato veliko variabilnost [87]. Zato so nekateri razdelitev večkrat ponovili, kot npr. v [70], kjer so razdelitev ponovili trikrat. Pristop prečnega preverjanja s pregibanjem je pogosto uporabljen na učni množici v postopku optimiziranja parametrov, kot je npr. število glavnih komponent pri metodah PCA ali PLS [76], število najbližjih sosedov pri metodi *k*-NN (*k*-NN, angl. *k*-Nearest Neighbours) [37] ipd. Kot pristop za končno vrednotenje uvrščanja NIRS podatkov je bila metoda prečnega preverjanja s pregibanjem redkeje uporabljena, npr. v [35] so uporabili prečno preverjanje s petimi pregibi, v [86] pa so uporabili prečno preverjanje z izpustitvijo ene enote (LOOCV, angl. Leave One Out Cross-Validation). S primeri podprto razpravo o izboru primernega števila pregibov pri metodi prečnega preverjanja lahko preberemo v [88].

Kljub množici obstoječih mer za vrednotenje uvrščanja (pregled mer najdemo v naslednjem razdelku 2.2.5, natančno predstavitev nekaterih mer pa v razdelku 3.3.3), se v študijah s področja NIRS večinoma uporabljajo preproste mere. V [5] so kot mero za vrednotenje uvrščanja uporabili število napačno uvrščenih enot, v [36, 37, 74] so uporabili napovedno točnost posameznih razredov in v [76] skupno napovedno točnost. Napovedna točnost (in prav tako število ali delež napačno uvrščenih enot) sta pri uvrščanju neuravnoteženih podatkov lahko zelo zavajajoči. Če so pri veliki stopnji neravnotežja vse enote pravilno uvrščene v večji razred, v manjši razred pa ni pravilno uvrščena nobena enota, potem je skupna napovedna točnost enaka deležu enot večjega razreda, kar je lahko v primeru velikega neravnotežja visoka vrednost, ki pa ne pove ničesar o delovanju modela uvrščanja.

#### 2.2.5 Pregled literature o merah za vrednotenje uvrščanja

Mera za vrednotenje napovedi uvrščanja je ocena delovanja klasifikatorja, ki je lahko izražena grafično ali numerično in jo lahko uporabimo za primerjanje klasifikatorjev med seboj. Problemi uvrščanja so lahko različni: pri nekaterih je pomembno predvsem pravilno uvrščanje v en razred, lahko so vsi razredi enako pomembni, ali pa je pomembna razlika med dejanskimi in napovedanimi vrednostmi [89]. Zato so tudi mere za vrednotenje uvrščanja različne in je pomembno, da za določen problem uvrščanja izberemo ustrezno mero.

Mere za vrednotenje napovedi uvrščanja so že dolgo predmet mnogih analiz in diskusij, vedno znova so predlagani tudi novi pristopi za ocenjevanje delovanja klasifikatorja. Mere

za vrednotenje napovedi uvrščanja so se sprva razvile za uvrščanje v dva razreda. O delovanju teh mer je bilo narejenih tudi več raziskav kot za mere za vrednotenje napovedi pri uvrščanju v več razredov. V Tabeli 2.2.1 so našteje mere za uvrščanje v dva razreda, ki bodo v nadaljevanju omenjene in jih ne moremo neposredno posplošiti za probleme uvrščanja v več razredov. V [90] najdemo pregled mer za uvrščanje v dva razreda, od mer za uvrščanje v več razredov pa avtorji obravnavajo le vzajemno informacijo (angl. mutual information). Mere za vrednotenje napovedi uvrščanja so za različne algoritme strojnega učenja primerjane tudi v članku [91]. V [92] je obravnavan vpliv neravnotežja na mere (PA, F, K, Krippendorfov  $\alpha$ , ROC in PR krivulje) pri uvrščanju v dva razreda. Vse mere razen ROC se izkažejo kot zelo občutljive na neravnotežje. Avtorji izrazijo celo možnost, da obravnavane mere povejo več o neravnotežju kot o delovanju modela uvrščanja. Pregled grafičnih mer za vrednotenje uvrščanja v dva razreda najdemo v [93].

Ime	Oznaka	Angleško ime	Vir
ROC krivulje	ROC	Receiver Operating characteristic Curves	[94]
površina pod ROC krivuljo	AUC	Area Under the ROC Curve	[95, 96]
Ginijev index	Gini	Gini index	[97]
površina pod Cohenovo krivuljo	AUK	Area Under the Cohen's curve	[98]
Krippendorfov alfa	$\alpha$	Krippendorf's alpha	[99, 100]
krivulja natančnost-občutljivost	PR krivulje	Precision-Recall curve	[101]

Tabela 2.2.1: Nekatere mere za vrednotenje napovedi uvrščanja pri uvrščanju v dva razreda, ki jih omenjamo in nimajo neposredne posplošitve za uvrščanje v več razredov.

Večina obstoječih mer za vrednotenje napovedi pri uvrščanju v več razredov se je razvila kot posplošitev mer za uvrščanje v dva razreda. Le redke so bile razvite s prvotnim namenom, ovrednotiti delovanje modelov za uvrščanje v več razredov. V Tabeli 2.2.2 so predstavljene mere za vrednotenje napovedi pri uvrščanju v več razredov. Večino mer iz prve skupine smo uporabili v analizi dejanskih podatkov in so natančneje predstavljene v razdelku 3.3.3.1.

Avtorji v študiji [136] so raziskovali lastnosti klasifikatorjev za uvrščanje v dva razreda, v več razredov, hierarhično uvrščanje in uvrščanje v več možnih razredov (angl. multi-labelled classification). Med seboj so primerjali tri skupine mer: mere, ki temeljijo na napovedni točnosti, na napovedni vrednosti ali na F-meri. Lastnosti mer so primerjali na podlagi invariantnih lastnosti osmih sprememb kontingenčne tabele: (1) zamenjava pozitivnega in negativnega razreda, (2) sprememba števila pravilno uvrščenih enot v negativni razred, (3) sprememba števila pravilno uvrščenih enot v pozitivni razred, (4) sprememba števila nepravilno uvrščenih enot v negativni razred, (5) sprememba števila nepravilno uvrščenih enot v pozitivni razred, (6) množenje kontingenčne tabele s skalarjem, (7) skalarno množenje stolpcev v kontingenčni tabeli in (8) skalarno množenje vrstic v kontingenčni tabeli. Izkazalo se je, da imajo mere iz iste skupine enake invariantne lastnosti.

Sigdel in Aygün [107] sta najprej predstavila pet zelenih lastnosti, ki naj bi jih mera za vrednotenje uvrščanja imela: (1) najvišja vrednost mere predstavlja klasifikator, ki uvrsti vse enote pravilno, (2) loči čim več različnih kontingenčnih tabel, (3) je invariantna na množenje kontingenčne tabele s skalarjem, (4) pri izračunu so vključuje vse vrednosti kontingenčne tabele, (5) uporabna je tako za uvrščanje v dva kot za uvrščanje v več razredov. Avtorja nato na desetih primerih uvrščanja v dva in desetih primerih uvrščanja v tri razrede preverita, katere od mer PA, K, MCC, CEN, F in Pacc imajo zelene lastnosti, in njihovo delovanje primerjata z delovanjem mere PA. Izkaže se, da so mere CEN, F-povprečje, MCC in K

Ime	Oznaka	Angleško ime	Vir
1. skupina: mere na podlagi praga uvrščanja			
napovedna točnost	PA	(Predictive) Accuracy	[102]
A-povprečje	A	macro-average Aritmetic	[103]
G-povprečje	G	macro-average Geometric, G-mean	[104]
Kappa	Kappa	Kappa	[98]
makro-povprečje napovedne vrednosti	MAP	Macro-Averaged Precision	[102]
F-povprečje	F	mean F-measure	[105]
korelacijski koeficient Matthew	MCC	Mathew's Correlation Coefficient, $k$ -category correlation coefficient $R_k$	[106]
verjetnostna napovedna točnost	Pacc	Probabilistic accuracy measure	[107]
po razredih uravnotežena napovedna točnost	CBA	Class Balance Accuracy	[108]
Jaccardov koeficient	J	Jaccard's coefficient or mean intersection over union	[109, 110]
indeks uspešnosti uvrščanja	CSI	Classification Success Index	[111]
kombinacija občutljivosti in napovedne točnosti	S-PA	Sensitivity-Accuracy approach	[112]
razvrstitvena entropija	CEN	Confusion Entropy	[113]
relativna informacija klasifikatorja	RCI	Relative Classifier Information	[114]
normalizirana vzajemna informacija	NMI	Normalized Mutual Information	[115, 116]
normalizirana prenesena informacija	H	normalized transmitted information	[117, 118]
vzajemna informacija	MI	Mutual Information	[119]
Youdenov index, informiranost	Y	Youden's index, informedness	[120, 121]
2. skupina: mere na podlagi rangiranja enot			
OAA AUC z neutženim povprečjem	AUNU	AUC of each class against the rest, using the uniform class distribution	[104]
OAA AUC z uteženim povprečjem glede na frekvenco razredov	AUNP	AUC of each class against the rest, using the a priori class distribution	[122]
OAo AUC z neutženim povprečjem	AU1U	AUC of each class against each other, using the uniform class distribution	[123]
OAo AUC z uteženim povprečjem glede na frekvenco razredov	AU1P	AUC of each class against each other, using the a priori class distribution	[104]
točkovana AUC	SAUC	scored AUC	[104, 124]
verjetnostna AUC	PAUC	Probabilistic AUC	[104]
prostornina pod ROC površino	VUS	Volume Under ROC Surface	[125]
3. skupina: mere na podlagi verjetnosti			
makro-povprečje verjetnosti	MAPR	Macro Average mean Probability Rate	[103]
povprečna verjetnost	MPR	Mean Probability Rate	[126]
povprečna absolutna napaka	MAE	Mean Absolute Error	
povprečna kvadratna napaka	MSE	Mean Squared Error, brier score	[127]
navzkrižna entropija	LogL	LogLoss, cross entropy	[128, 129]
kalibracijska izguba	CalL	Calibration Loss	[130]
kalibracija po delih	CalB	Calibration by Bins	[131]
entropijska napovedna točnost	EMA	Entropy-Modulated Accuracy	[132]
normaliziran informacijski prenos	NIT	Normalized Information Transfer	[133]
informacijska variabilnost	VI	Variation of Information	[134]
4. skupina: grafične mere			
entropijski trikotnik	ET	De Finetti Entropy Triangle	[135]

Tabela 2.2.2: Mere za vrednotenje uvrščanja v več razredov.



manj skladne z mero PA kot mera Pacc. Največ različnih kontingenčnih tabel zazna mera CEN, nato sledijo Pacc, F-povprečje in MCC, daleč najmanj različnih kontingenčnih tabel ločita meri PA in K. Mera Pacc je tudi najmanj občutljiva na množenje matrike s skalarjem, medtem ko med najbolj občutljive sodita K in PA.

Vpliv neravnotežja na mere za vrednotenje napovedi pri uvrščanju v več razredov so raziskovali v [104] in [137]. Ferri in sodelavci [104] so mere razdelili v tri skupine:

1. mere, ki ocenjujejo klasifikator na podlagi praga uvrščanja, so tiste, ki merijo število napačno in pravilno uvrščenih enot (PA, A, G, F, K ...);
2. mere, ki ocenjujejo klasifikator na podlagi rangiranja enot, so tiste, ki merijo predvsem, kako dobro klasifikator loči med razredi (mere izpeljane iz AUC);
3. mere, ki ocenjujejo klasifikator na podlagi verjetnosti, ne merijo le, koliko enot je bilo pravilno ali napačno uvrščenih, ampak merijo tudi, s kakšno verjetnostjo se je to zgodilo (LogL, MSE ...).

Ob razvrstitvi mer v opisane skupine so v [104] v eksperimentalni študiji s 30 podatkovnimi bazami raziskali tudi, kako skladni so rezultati 18 mer za vrednotenje napovedi uvrščanja (mere so bile iz vseh treh skupin). Podatkovne baze so vsebovale od 2 do 19 razredov, pri čemer so bile nekatere uravnotežene, druge pa ne. Ugotovili so, da tudi mere, ki sicer sodijo v isto skupino, merijo različne stvari in še posebej za neuravnotežene podatke in podatke z več razredi dajejo neskladne rezultate. Zelo podobno delujejo edino mere na podlagi AUC, rezultati verjetnostnih mer pa so med seboj najmanj usklajeni.

V članku [137] so primerjali sedem mer za vrednotenje uvrščanja (G, A, F in štiri posplošitve mere AUC za problem uvrščanja v več razredov) na petih podatkovnih bazah, ki so imele več kot dva razreda in so bile neuravnotežene. Model uvrščanja so zgradili z umetnimi nevronskimi mrežami tipa večnivojski perceptron (angl. multilayer perceptron), uporabili so tudi pet pristopov za zmanjšanje vpliva neravnotežja. Ugotovitve, ki so iz študije sledile, so bile podobne kot v [104], in sicer: mere za vrednotenje napovedi uvrščanja so med seboj usklajene, vendar ni vedno dovolj uporabiti skupne mere za primerjavo klasifikatorjev in algoritmov. V nekaterih primerih so ob visokih vrednostih mer za vrednotenje skupne napovedi uvrščanja opazili nizke vrednosti mer manjših razredov, medtem ko so bile v drugih primerih mere vrednotenja skupne napovedi s posameznimi merami manjših razredov usklajene. Avtorji zaključijo, da ta različnost rezultatov kaže, da mere za vrednotenje skupne napovedi uvrščanja ne odražajo vedno dejanskega izboljšanja ali poslabšanja uvrščanja, čeprav se je le-to morda zgodilo.

Mere CEN, RCI, NMI, H, MI, LogL, EMA, NIT in VI izhajajo iz informacijske teorije. Podobnost mere CEN z MCC so raziskovali in potrdili v [138]. Valverde-Albacete in Peláez-Moreno sta primerjala mere PA, MCC, CEN, EMA, NIT in grafično mero ET. Ugotovila sta, da MCC deluje podobno kot PA in ima zato tudi podobne slabosti (ne deluje za neuravnotežene podatke). CEN deluje sicer nekoliko bolje, vendar je še vedno močno pristranska v prid večjemu razredu. Za naloge uvrščanja, kjer ni pomembno le, koliko enot je uvrščenih pravilno in koliko napačno, avtorja priporočata verjetnostni meri EMA in NIT.

## 2.3 Generiranje NIRS podatkov v dosedanjih študijah

Več avtorjev je poskusilo umetno generirati NIRS podatke, da bi z njimi v simulacijah lahko primerjali delovanje statističnih metod. Tan in Brown [55] sta generirala podatke s pomočjo Gaussovih krivulj, da bi preverila delovanje nove metode predobdelave spektrov, ki temelji na odstranitvi trenda v ozadju (angl. non-constant background correction). Simulirala sta Gaussovi krivulji z različnima širinama, pri čemer je širša predstavljala trend v ozadju, ožja pa analitični signal. Podoben pristop je bil uporabljen tudi v študiji [46]. V Ge in sodelavci [47] so bili umetni NIRS podatki prav tako uporabljeni za testiranje delovanja nove metode predobdelave spektrov, ki je bila nadalje uporabljena v regresijskih modelih. Spektri odboja so bili generirani na podlagi šestih Gaussovih krivulj z različnimi širinami in višinami. Pozornost pri generiranju podatkov je bila posvečena predvsem vrhovom Gaussovih krivulj, ki so bili določeni tako, da sta se po dva vrhova popolnoma prekrivala, delno prekrivala ali pa se nista prekrivala.

V večini študij, ki vključujejo simulacijo NIRS podatkov, so bili umetni podatki generirani za simuliranje NIRS podatkov kemijskih zmesi [49–54] (kemijsko zmes dobimo z mešanjem več čistih snovi, pri čemer kemijska reakcija ne poteče). Umetni podatki so bili nato uporabljeni za primerjavo delovanja različnih regresijskih pristopov s ciljem, oceniti koncentracijo posameznih sestavin kemijske zmesi. NIRS podatki kemijskih zmesi so bili generirani po Beer-Lambertovem zakonu o aditivnosti absorbanč: izračunane so bile linearne kombinacije spektrov čistih snovi v zmesi (produkt matrike koncentracij in spektrov čistih snovi), na koncu je bil dodan šum. Wentzel in Montono [53] sta generirala spektre čistih snovi tako, da sta večkrat zaporedoma uporabila metodo premikajočega povprečja na podatkih, pridobljenih naključno iz standardne normalne porazdelitve. Končni podatki so imeli sicer sprejemljive lastnosti, vendar sta avtorja izrazila negotovost o tem, koliko se takšen način generiranja podatkov zares približa realnosti. V študiji [54] je bil spekter za vsako čisto snov generiran z eno Gaussovo krivuljo. Podatkom je bil nato dodan šum v obliki zapletenih kovariančnih struktur. Podobno sta Hemmateenejad in Karimi [52] generirala spektre čistih snovi z enim ali dvema Gaussovima krivuljama. Generirala sta tudi 20 trendov, ki sta jih naključno prištela generiranim spektrom zmesi in na koncu dodala normalno porazdeljene napake. Podoben pristop je bil v študijah [49–51] uporabljen za proučevanje metod za izbor spremenljivk v regresijskih modelih.

V literaturi lahko najdemo le malo primerov, kjer bi generirali NIRS podatke z namenom proučevanja metod uvrščanja. Wang in sodelavci [48] so testirali novo metodo za intervalni izbor spremenljivk v kombinaciji z metodo SVM za uvrščanje dveh skupin materialov. Pri tem so uporabili tako realne kot tudi umetno generirane podatke. Spektralni podatki so bili najprej generirani kot naključna števila iz normalne porazdelitve. Polovici generiranih spektrov je bil dodan trend v obliki širših Gaussovih krivulj. Neodvisno so nato izbrali polovico spektrov in jim na točno določenem mestu dodali signal v obliki ožje Gaussove krivulje.

## Poglavje 3

# Metode

Poglavje Metode je razdeljeno na pet razdelkov. V prvem (razdelek 3.1) so predstavljeni dejanski podatki 45 vrst polimerov in strnjen opis obdelave teh podatkov. V drugem in tretjem razdelku (razdelek 3.2 in razdelek 3.3) so natančno opisane metode, ki smo jih uporabili pri gradnji modelov uvrščanja tako na realnih kot na simuliranih podatkih. V razdelku 3.4 so opisane metode umetnega generiranja podatkov, s katerimi smo generirali podatke, ki smo jih kasneje uporabili v simulacijski študiji. Nastavitve simulacij pa so predstavljene v razdelku 3.5.

### 3.1 Dejanski podatki plastik

V vzorčni zbirki zgodovinskih polimerov je bilo 534 predmetov (enote), ki so pripadali 45 različnim vrstam polimerov (razredi). Podatki so bili močno neuravnoteženi, saj je najmanjši razred vseboval 3, največji razred pa 59 enot (Tabela 3.1.1). Pravilo za uvrščanje smo zgradili na podlagi odboja svetlobe v NIR območju. Moč odboja smo zmerili s spektrometrom Labspec NIR256-2.5 Near-Infrared Spectrometer Ocean Optics (Ocean optics, USA) z razponom valovnih dolžin med 900 in 2500 nm in vzorčnim intervalom med 6 in 7 nm. Zaradi velikega šuma na obeh robovih opazovanega NIR območja smo vzeli le meritve pri valovnih dolžinah med 1058 in 2294 nm in tako dobili 191 spremenljivk. Podatki so last podjetja Morana-RTD, d. o. o. in so bili zbrani v okviru Sedmega okvirnega programa EU Popart [139]. Slike vseh spektrov, obdelanih s SNV predobdelavo, so v dodatku A.

Za izgradnjo učinkovitega pravila za uvrščanje plastik v razrede glede na vrsto plastike smo preizkusili 4 metode predobdelave spektrov, 4 metode zmanjšanja dimenzije podatkov in 8 metod uvrščanja. Celotno analizo smo izvedli s programskim paketom **R** [140].

Za predobdelavo spektrov smo uporabili naslednje metode (natančneje predstavljene v razdelku 3.2):

- Brez predobdelave.
- Standardna normalna vektorska transformacija (SNV, angl. *standard normal variate*), ki smo jo izračunali s funkcijo *scale* iz osnovnega paketa funkcij programa **R**.
- Kvantilna normalizacija, za katero smo uporabili funkcijo *normalize.quantiles.robust* iz paketa *preprocessCore* [141].

Oznaka	Polimer	Učna množica	Testna množica	Vsota
1	Poly(butylene terephthalate)	2	1	3
2	Cellophane	2	1	3
3	Polyamide	2	2	4
4	Silk	2	2	4
5	Acrylonitrile-butadiene-styrene	3	2	5
6	Polyvinylchloride/polyvinylacetate	3	2	5
7	Casein formaldehyde	3	2	5
8	Melamine formaldehyde resin	3	2	5
9	Polysulfone	3	2	5
10	Poly(vinyl acetate)	3	2	5
11	Polyisoprene	3	2	5
12	Poly(phenylene oxide)	3	2	5
13	Poly(phenylene sulfide)	3	2	5
14	Phenol-formaldehyde	3	2	5
15	Polyamide	3	2	5
16	Polycarbonate	4	2	6
17	Polytetrafluoroethylene	4	2	6
18	Leather	4	2	6
19	Cellulose acetate propionate	4	3	7
20	Styrene-acrylonitrile	5	3	8
21	Natural rubber	5	3	8
22	Poly(methyl methacrylate)	5	3	8
23	Cellulose acetate butyrate	6	3	9
24	Cellulose acetate	6	3	9
25	Styrene-butadiene	6	4	10
26	Poly(propylene oxide)	6	4	10
27	Ethylene-vinyl acetate	6	4	10
28	Polyurethane (ester) (PUR)	6	4	10
29	Vinyl chloride-vinyl acetate-vinyl alcohol	6	4	10
30	Amber	6	4	10
31	Urea-formaldehyde resin	6	4	10
32	Epoxy resin	6	4	10
33	Polyurethane (ether) (PUR)	7	4	11
34	Polyoxymethylene	7	4	11
35	Polydimethylsiloxane	8	5	13
36	Bones, Teeth, Horns, Tortoiseshell, Ivory	8	5	13
37	Poly(ethylene terephthalate)	9	5	14
38	Polyamide	10	5	15
39	Urethane elastomer thermoplastic	12	7	19
40	Polyethylene	17	9	26
41	Polypropylene	18	9	27
42	Cotton/linen	18	9	27
43	Cellulose nitrate	26	14	40
44	Poly(vinyl chloride)	28	15	43
45	Polystyrene	39	20	59
Vsota		339	195	534

Tabela 3.1.1: Vrste polimerov, število enot v učni in testni množici ter skupno število enot.

- Prvi odvod izračunan po metodi Savitzky-Golay s funkcijo *sgolayfilt* iz paketa *signal* [142] s širino okna 7 in polinomom prve stopnje.

Da bi preizkusili, ali zmanjšanje dimenzije podatkov (podroben opis metod je v razdelku 3.2) vpliva na napovedno točnost, smo v izgradnjo modelov uvrščanja vključili:

- vseh 191 spremenljivk;

- izbor 50 spremenljivk z največjo varianco;
- izbor 50 spremenljivk z največjo  $F$  statistiko;
- glavne komponente, izračunane po metodi PCA, ki so pojasnile 99 % celotne variabilnosti podatkov, pri tem smo uporabili funkcijo *prcomp* iz paketa *stats* [143].

Uporabljene metode uvrščanja, ki so natančnejše predstavljene v razdelku 3.3.1 so:

- linearna diskriminantna analiza (LDA) (funkcija *lda* iz paketa *MASS* [144] s privzetimi nastavitvami);
- metode najbližjega soseda ( $k$ -NN) z 1, 3 in 5 sosedi (funkcija *knn* iz paketa *FNN* [145] s privzetimi nastavitvami);
- klasifikacijska in regresijska drevesa (CART) (funkcija *rpart* iz paketa *rpart* [146], če je bilo število enot v vozlišču 2 ali več, se je nadaljnja delitev še lahko zgodila, za vse ostale argumente funkcije *rpart* pa smo obdržali privzete nastavitve: apriorni delež je bil enak deležu enot v učni množici, matrika izgube je bila 1 in kriterijska funkcija je bila Ginijev indeks);
- metoda podpornih vektorjev (R-SVM) z radialno jedrno funkcijo (funkcija *svm* iz paketa *e1071* [147]);
- metoda podpornih vektorjev (L-SVM) z linearno jedrno funkcijo (funkcija *svm* iz paketa *e1071* [147]) in prilagojenim pragom za uvrščanje na način, da smo na vsakem koraku metode vsak proti vsakem za prag uvrščanja določili delež enot 1. razreda.

Z različnimi kombinacijami vseh naštetih metod smo ovrednotili 112 različnih modelov uvrščanja.

Za izbiro optimalnega modela smo uporabili dva postopka: 500 krat ponovljeno razdelitev na učno in testno množico in 100 krat ponovljeno prečno preverjanje z 10 pregibi (glej razdelek 3.3.4). Pri 500 krat ponovljeni razdelitvi smo v vsakem od 500 ponovitev podatke razdelili na učno (339) in testno (195) množico tako, da so bile tako v učni kot testni množici zastopane enote iz vsakega razreda (Tabela 3.1.1). Na učni množici smo zgradili model uvrščanja in ga nato na testni množici ovrednotili, tako da smo izračunali mere za vrednotenje napovedi uvrščanja, opisane v razdelku 3.3.3. Te vrednosti smo nato povprečili po vseh 500 ponovitvah. Pri 100 krat ponovljenem prečnem preverjanju z 10 pregibi smo v vsaki od 100 ponovitev podatke razdelili na 10 disjunktnih podmnožic tako, da so bili razredi med podmnožice porazdeljeni, kar se da enakomerno. Tudi tukaj smo pri vsaki od 100 ponovitev izračunali mere za vrednotenje napovedi uvrščanja ter na koncu te vrednosti povprečili po vseh 100 ponovitvah.

Preizkusili smo tudi delovanje dveh metod za zmanjšanje vpliva neravnotežja: večkratno zmanjšanje večjega razreda (MDS, angl. multiple downsizing) in vsak proti vsakem (OAO, angl. one-against-one). Metodo MDS smo preizkusili pri modelih, zgrajenih z metodami LDA, CART, R-SVM in L-SVM v kombinaciji z vsako od štirih omenjenih predobdelav in metod zmanjšanja dimenzije podatkov, pri tem smo iz osnovne učne množice vzorčili 100 novih uravnoteženih učnih množic. Metodo OAO smo preizkusili le pri modelih, zgrajenih z metodami LDA, CART in R-SVM, a le v kombinacijami s SNV predobdelavo, ki se je v večini primerov izkazala za najuspešnejšo. Modele, ki smo jih dobili z uporabo metod MDS in OAO, smo ovrednotili le z metodo 100 krat ponovljenega prečnega preverjanja z 10 pregibi.

## 3.2 Metode predobdelave in zmanjšanja dimenzije podatkov

Predobdelava spektra je matematičen postopek, katerega namen je iz spektralnih meritev izključiti motečo variabilnost, ki je pogosto vzrok lastnosti merjenih snovi, kot so oblika, velikost delcev, barva, osvetlitev prostora ipd. V literaturi je bilo predstavljenih že veliko metod predobdelave [79]. Metoda, ki bi uspešno delovala za vse podatke, ne obstaja, zato primerno metodo določimo za vsak problem obdelave posebej. V doktorskem delu smo se osredotočili na tri metode predobdelave, ki so predstavljene v nadaljevanju: standardna normalna vektorska transformacija (SNV, angl. standard normal variate), kvantilna normalizacija in prvi odvod, izračunan po metodi Savitzky-Golay.

Z metodami zmanjšanja dimenzije podatkov zmanjšamo število spremenljivk, na podlagi katerih nato zgradimo model uvrščanja. Postopek izgradnje modelov uvrščanja je tako časovno učinkovitejši, modeli uvrščanja pa so manj kompleksni. S tem se izognemo težavi preprileganja, ki je še posebej izrazita pri visoko razsežnih koreliranih podatkih, kot so NIRS podatki. Z zmanjšanjem dimenzije podatkov omogočimo tudi uporabo tistih metod uvrščanja, ki jih v primeru večjega števila spremenljivk kot enot ni mogoče uporabiti. V nadaljevanju so natančno predstavljene metode zmanjšanja dimenzije podatkov, ki smo jih v doktorskem delu uporabili.

Za lažje opisovanje metod bomo z  $N$  označili število enot, s  $p$  število spremenljivk in z  $x_{ij}$  vrednost  $i$ -te enote na  $j$ -ti spremenljivki. Oznaka  $x_i$  predstavlja vrednosti  $i$ -te enote pri vseh spremenljivkah, kar imenujemo tudi spekter  $i$ -te enote. Vrednosti  $j$ -te spremenljivke pri vseh enotah označimo z  $x_{.j}$ .

### 3.2.1 Standardna normalna vektorska transformacija

Standardna normalna vektorska transformacija (SNV, angl. Standard Normal Variate) [148] je standardizacija enot in jo za vsak spekter  $x_i$  izračunamo kot

$$x_i^{stand} = \frac{x_i - \bar{x}_i}{s_i},$$

kjer je  $\bar{x}_i$  povprečje in  $s_i$  standardni odklon spektra  $i$ -te enote  $x_i$ . Za izračun SNV predobdelave smo uporabljali funkcijo *scale* iz osnovnega paketa funkcij programa **R**.

### 3.2.2 Kvantilna normalizacija

S kvantilno normalizacijo [149] poenotimo moči spektrov med enotami. V ta namen meritve spektrov za vsako enoto posebej rangiramo in nato vsem vrednostim z istim rangom  $q$  priredimo vrednost  $x^q$ . Za  $x^q$  smo v naših analizah vzeli povprečje vseh vrednosti z rangom  $q$ .

Kvantilno normalizacijo smo izvedli s funkcijo *normalize.quantiles.robust* iz paketa *preprocessCore* [141].

### 3.2.3 Prvi odvod z metodo Savitzky-Golay

Z uporabo prvega odvoda iz spektralnih podatkov učinkovito odstranimo variabilnost v osnovnih močeh med enotami, ohrani pa se povezava spektra s kemičnimi lastnostmi. S prvim odvodom se v podatkih poudarijo tudi zelo majhni premiki, ki pogosto predstavljajo šum. Zato je spekter, obdelan s prvim odvodom, uporaben samo v kombinaciji z glajenjem, ki mora biti takšno, da zgladi šum in hkrati ohrani vso pomembno informacijo.

Prvi odvod v  $j$ -ti točki spektra  $x_i$  z metodo Savitzky-Golay [150] izračunamo tako, da na simetričnem oknu širine  $w$  okoli  $j$ -te točke po metodi najmanjših kvadratov ocenimo koeficiente polinoma stopnje  $r$ . Ko so koeficienti polinoma določeni, ga analitično odvajamo in izračunamo vrednost odvoda v točki  $x_{ij}$ . Ta postopek ponovimo za vsako točko spektra  $x_i$ . Na robovih spektra okno okoli točke, v kateri računamo odvod, ni simetrično.

Za izračun prvega odvoda smo uporabili funkcijo *sgolayfilt* iz paketa *signal* [142] z oknom širine 7 in stopnjo polinoma 1. V začetnih izračunih smo na manjši množici podatkov preizkusili tudi višje stopnje polinomov, ki pa niso dali boljših rezultatov, zato smo se odločili, da v nadaljevanju uporabimo polinom 1. stopnje, ki je tudi za izračun najmanj zahteven. Prav tako smo preizkusili različne širine okna (5, 7 in 9), pri čemer se je okno širine 7 izkazalo za najuspešnejše. Rezultati teh izračunov v doktorskem delu niso prikazani.

### 3.2.4 Izbor spremenljivk z največjo varianco

Pri izboru spremenljivk glede na največjo varianco izračunamo vzorčno varianco vseh spremenljivk. Za nadaljnjo analizo so izbrane spremenljivke z največjo vzorčno varianco. V naših izračunih smo izbrali 50 spremenljivk z največjo vzorčno varianco.

### 3.2.5 Izbor spremenljivk z največjo $F$ -statistiko

$F$ -statistika je največja za tiste spremenljivke, ki najbolj ločijo skupine. Za  $j$ -to spremenljivko jo izračunamo kot

$$F_j = \frac{\text{variabilnost med skupinami}_j}{\text{variabilnost znotraj skupin}_j}$$

$$\text{variabilnost med skupinami}_j = \frac{1}{K-1} \sum_{k=1}^K n_k \left( \bar{x}_{\cdot j}^k - \bar{x}_{\cdot j} \right)^2 \quad (3.1)$$

$$\text{variabilnost znotraj skupin}_j = \frac{1}{N-K} \sum_{k=1}^K \sum_{i=1}^{n_k} \left( x_{ij} - \bar{x}_{\cdot j}^k \right)^2, \quad (3.2)$$

kjer je  $\bar{x}_{\cdot j}^k$  povprečje  $j$ -te spremenljivke v  $k$ -ti skupini,  $\bar{x}_{\cdot j}$  povprečje  $j$ -te spremenljivke,  $K$  število skupin in  $n_k$  število enot v  $k$ -ti skupini. V naših izračunih smo izbrali 50 spremenljivk z največjo  $F$  statistiko.

### 3.2.6 Metoda glavnih komponent (PCA)

Glavne komponente [151] so linearne kombinacije merjenih spremenljivk, ki jih poiščemo tako, da imajo le-te največjo varianco in so med seboj pravokotne. Glavne komponente si sledijo od prve do zadnje po padajoči varianci. Prva komponenta tako poteka v smeri največje variabilnosti podatkov, zato lahko sklepamo, da nosi največ informacije, ki jo lahko z linearno kombinacijo dobimo iz merjenih spremenljivk. Za izračun PCA smo uporabili funkcijo *prcomp* iz osnovnega paketa funkcij programa **R** s privzetimi nastavitvami. V modele uvrščanja smo vključili takšno število glavnih komponent, da so skupaj pokrile vsaj 99 % skupne variabilnosti podatkov.

## 3.3 Metode uvrščanja in vrednotenja modelov uvrščanja

Za lažje opisovanje metod uvrščanja in vrednotenja modelov uvrščanja vpeljimo najprej nekaj oznak.

Naj bo  $N$  število enot,  $p$  število spremenljivk,  $K$  število razredov in  $n_k$  število enot v  $k$ -tem razredu. Vrednost  $i$ -te enote na  $j$ -ti spremenljivki označimo z  $x_{ij}$ . Z  $x_i$  označimo vrednosti  $i$ -te enote pri vseh spremenljivkah, kar imenujemo tudi spekter  $i$ -te enote. Z  $x_{.j}$  pa označimo vrednosti  $j$ -te spremenljivke pri vseh enotah.

Klasifikator ali model uvrščanja je funkcija, ki vsaki enoti na podlagi vrednosti spremenljivk priredi natančno en razred  $k \in \{1, 2, \dots, K\}$ . Kontingenčna tabela (Tabela 3.3.1) delovanja klasifikatorja  $C$  je  $K \times K$  matrika, kjer poljuben element matrike  $c_{ij}$  predstavlja število enot iz razreda  $i$ , ki jim je klasifikator priredil razred  $j$ . Z  $n'_k$  označimo število enot, ki jih je klasifikator uvrstil v  $k$ -ti razred.

		Napovedan razred				Skupaj
		$R_1$	$R_2$	$\dots$	$R_K$	
Dejanski razred	$R_1$	$c_{11}$	$c_{12}$	$\dots$	$c_{1K}$	$n_1$
	$R_2$	$c_{21}$	$c_{22}$	$\dots$	$c_{2K}$	$n_2$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$R_K$	$c_{K1}$	$c_{K2}$	$\dots$	$c_{KK}$	$n_K$
Skupaj		$n'_1$	$n'_2$	$\dots$	$n'_K$	$N$

Tabela 3.3.1: Kontingenčna tabela delovanja klasifikatorja

### 3.3.1 Metode uvrščanja

#### 3.3.1.1 Linearna diskriminantna analiza

Pri diskriminantni analizi [152] iščemo takšne linearne kombinacije merjenih spremenljivk (diskriminantne spremenljivke), ki najbolj ločijo v naprej določene skupine. Diskriminantne spremenljivke poiščemo tako, da je razmerje med variabilnostjo med skupinami in variabilnostjo znotraj skupin največje.



Centroid  $k$ -te skupine imenujemo vektor povprečij diskriminantnih spremenljivk v  $k$ -ti skupini. Ko želimo za novo enoto določiti pripadnost razredu, izračunamo evklidske razdalje med vrednostmi nove enote na diskriminantnih spremenljivkah in centriidi vseh skupin. Nova enota je nato uvrščena v tisto skupino, kjer je ta razdalja najmanjša. Kadar za enote znotraj skupine predpostavimo normalno porazdelitev in enako varianco med skupinami, govorimo o linearni diskriminantni analizi (LDA, angl. Linear Discriminant Analysis). Za izračun LDA smo uporabili funkcijo *lda* iz paketa *MASS* [144] s privzetimi nastavitvami.

### 3.3.1.2 Metode najbližjega soseda

Pri metodi  $k$  najbližjih sosedov (k-NN, angl.  $k$ -Nearest Neighbours) [153, 154] izračunamo razdalje med novo enoto in enotami v učni množici. Nato poiščemo  $k$  enot iz učne množice, ki so novi enoti najbližje – to je  $k$  najbližjih sosedov. Novo enoto potem uvrstimo v tisti razred, v katerem je večina izmed  $k$  najbližjih sosedov.

Za uvrščanje z metodo k-NN smo uporabili funkcijo *knn* iz paketa *FNN* [145] s številom razredov  $k \in \{1, 3, 5\}$  in privzetimi nastavitvami, pri čemer je bila izračunana razdalja med enotami evklidska.

### 3.3.1.3 Odločitvena drevesa

Metoda odločitvenih dreves deluje tako, da na podlagi učne množice razdeli prostor spremenljivk na disjunktne podmnožice (končna vozlišča ali listi). Enotam iz lista  $l$  priredimo isti razred, in sicer tisti, ki v učni množici na listu  $l$  prevladuje. Eden izmed algoritmov za izgradnjo odločitvenih dreves je CART (CART, angl. Classification And Regression Trees) [155], ki smo ga uporabili v naših izračunih.

Algoritem CART je binarni rekurzivni postopek, kar pomeni, da gradi drevo postopoma in na vsakem koraku naredi eno samo delitev na dva dela, ki je v tistem koraku najboljša glede na kriterijsko funkcijo. V prvem koraku algoritma izberemo spremenljivko  $j$  in vrednost  $a$  tako, da razdelitev prostora na podmnožici, kjer je  $x_{.j} < a$  ali  $x_{.j} \geq a$ , najbolj izboljša kriterijsko funkcijo. V naslednjem koraku izberemo eno od podmnožic iz prejšnjega koraka, eno spremenljivko in eno vrednost, ki najbolj izboljšajo kriterijsko funkcijo. Postopek rekurzivno ponavljamo, dokler ni izpolnjen ustavitveni pogoj. Na tak način razbijemo prostor spremenljivk na  $M$  večdimenzionalnih pravokotnikov. Podmnožicam, ki jih dobimo v vmesnih korakih, rečemo vozlišča, končnim podmnožicam pa končna vozlišča ali listi. Kot kriterijska funkcija je pogosto uporabljen Ginijev indeks

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

kjer  $\hat{p}_{mk}$  predstavlja delež enot iz razreda  $k$  v  $m$ -ti podmnožici. Ginijev indeks meri stopnjo čistosti vozlišča, pri čemer manjše vrednosti pomenijo večjo čistost. Vozlišče je čisto, če so v njem vse enote iz istega razreda.

Za izgradnjo modelov uvrščanja po metodi CART smo uporabili funkcijo *rpart* iz paketa *rpart* [146]. Če je bilo število enot v vozlišču 2 ali več, se je nadaljnja delitev še lahko zgodila. Za vse ostale argumente funkcije *rpart* smo obdržali privzete nastavitve: apriorni delež je

bil enak deležu enot v učni množici, matrika izgube je bila 1 in kriterijska funkcija je bila Ginijev indeks.

### 3.3.1.4 Metoda podpornih vektorjev

Pri metodi podpornih vektorjev (SVM, angl. Support Vector Machines) [156] iščemo v prostoru s  $p$  spremenljivkami hiperravnino dimenzije  $p - 1$ , ki bi ločila dve skupini učnih podatkov (vektorjev dimenzije  $p$ ). Če sta skupini linearno ločljivi, potem ne samo da takšna hiperravnina obstaja, ampak jih obstaja celo več. Za hiperravnino, ki najbolj loči podatke med seboj, izberemo tisto, ki je najbolj oddaljena od najbližjih točk v obeh skupinah. Okoli hiperravnine tako dobimo razmejitveni pas, v katerem ni nobene učne enote. Ta pas predstavlja mejo med skupinama. Širši kot je, bolj zanesljivo je uvrščanje. Novo enoto nato uvrstimo v skupino glede na to, na kateri strani hiperravnine leži.

V primeru, ko skupini nista ločljivi oz. je med njima le ozka meja, določimo tako imenovano mehko mejo, kjer nekaj učnih vektorjev "žrtvujemo", da ležijo znotraj razmejitvenega pasu ali celo na napačni strani hiperravnine, z namenom izboljšati zanesljivost uvrščanja za nove enote. Izkaže se, da je enačba hiperravnine odvisna le od teh vektorjev (enot) in jih zato imenujemo podporni vektorji.

Skupini sta lahko ločljivi, ampak ta ločljivost ni nujno linearna. To pomeni, da v prostoru obstoječih spremenljivk dimenzije  $p$  ne obstaja hiperravnina, ki bi skupini ločila. Če dovolj in na pravi način povečamo dimenzijo prostora, se izkaže, da sta skupini v tem novem prostoru linearno ločljivi. Povečanje dimenzije prostora in izračun metode podpornih vektorjev je lahko računsko zelo zapleten in pogosto tudi neizvedljiv postopek. Pri metodi podpornih vektorjev se povečanje dimenzije prostora reducira na uporabo tako imenovanih jedrnih funkcij. Jedrne funkcije ( $K$ , angl. Kernel) so funkcije, ki merijo podobnost med spremenljivkami in jih v izračunu metode podpornih vektorjev uporabimo kot posplošitev skalarnega produkta, ki igra v izračunu modela z metodo SVM ključno vlogo. Funkcijo običajnega skalarnega produkta med enotami

$$K(x_i, x_{i'}) = \sum_{j=1}^p x_{ij} x_{i'j}$$

imenujemo linearna jedrna funkcija, ki za oceno podobnost med enotama uporabi Pearsonovo korelacijo. Pogosto je uporabljena tudi radialna jedrna funkcija

$$K(x_i, x_{i'}) = \exp \left( -\gamma \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right); \gamma > 0,$$

ki da zelo majhne vrednosti, ko so enote daleč narazen, pri čemer je uporabljena evklidska razdalja.

Za uvrščanje z metodo SVM smo uporabili funkcijo *svm* iz paketa *e1071* [147] z radialno in linearno jedrno funkcijo in privzetimi nastavitvami. Ker metoda SVM deluje le za uvrščanje v dva razreda, funkcija *svm* za uvrščanje v več razredov uporablja pristop vsak proti vsakem, ki je natančneje opisan v razdelku 3.3.2.1.

### 3.3.2 Pristopi za zmanjšanje neravnotežja pri uvrščanju v več razredov

#### 3.3.2.1 Pristopa vsak proti vsakem in vsak proti vsem

Pri uvrščanju v več razredov lahko uporabimo eno od metod vsak proti vsakem (OAO, angl. One-Against-One) ali vsak proti vsem (OAA, angl. One-Against-All). Pri metodi OAO za vsak par razredov zgradimo model uvrščanja, to je skupaj  $\binom{K}{2}$  klasifikacijskih modelov, kjer je  $K$  število razredov. Novo enoto nato napovemo z vsemi  $\binom{K}{2}$  modeli. Končno pripadnost razredu določimo z večinskim glasovanjem (angl. majority voting): enota je uvrščena v razred, ki se je med  $\binom{K}{2}$  napovedmi pojavil največkrat. Pri metodi OAA pa gradimo modele uvrščanja tako, da en razred predstavlja prvo skupino, vsi ostali razredi predstavljajo drugo skupino. Na tak način zgradimo  $K$  modelov uvrščanja. Novo enoto napovemo z vsemi  $K$  modeli in jo na koncu uvrstimo v tisti razred, v katerega je bila napovedana z največjo verjetnostjo. Slabost pristopa OAA je, da so binarni modeli uvrščanja, ki jih gradimo, zelo neuravnoteženi, še posebej v primeru velikega števila razredov, zato smo pri našem delu preizkusili le metodo OAO.

#### 3.3.2.2 Večkratno zmanjšanje večjega razreda

Metoda večkratnega zmanjšanja večjega razreda (MDS, angl. Multiple Downsizing) je različica metode asimetrični *bagging* [157]. Iz učne množice z vzorčenjem brez ponavljanja tvorimo vrsto novih učnih množic tako, da iz večjih razredov izberemo toliko enot, kot jih je v najmanjšem razredu. Na vsaki novi učni množici zgradimo klasifikator. Novo enoto nato napovemo z vsemi klasifikatorji, končno pripadnost razredu pa določimo z večinskim glasovanjem, to pomeni, da enoto uvrstimo v tisti razred, v katerega bi jo uvrstila večina klasifikatorjev. V izračunih v doktorskem delu smo na opisan način vzorčili 100 novih učnih množic.

#### 3.3.2.3 Prilagoditev praga za uvrščanje

Pri klasifikatorjih, ki uvrščajo v dva razreda in ocenijo verjetnost, s katero je enota uvrščena v določen razred, moramo določiti prag za uvrščanje. Naj bo  $p_1(x)$  posteriorna verjetnost, ki jo vrne klasifikator, da enota  $x$  pripada 1. razredu, in naj bo  $P$  prag za uvrščanje, potem pripadnost razredu  $c(x)$  za enoto  $x$  določimo tako:

$$c(x) = \begin{cases} 1; & p_1(x) > P \\ 2; & p_1(x) < P \\ \text{rand}(\{1,2\}); & p_1(x) = P, \end{cases}$$

kjer funkcija  $\text{rand}(A)$  vrne naključno število iz množice  $A$ . Prag za uvrščanje  $P$  je običajno enak 0,5. V primeru neuravnotežene učne množice lahko vpliv neravnotežja zmanjšamo tako, da prag  $P$  ustrezno prilagodimo glede na velikost razredov, pri čemer je naravna izbira za prag  $P$  delež enot 1. razreda v učni množici.

### 3.3.3 Mere za vrednotenje uvrščanja

Glede na vrednost, ki jo klasifikator vrne, ločimo verjetnostne (probabilistične) in neverjetnostne klasifikatorje. Verjetnostni klasifikatorji ocenijo verjetnost, s katero lahko novo enoto uvrstimo v določen razred, medtem ko neverjetnostni vrnejo le pripadnost razredu. Če pri verjetnostnih klasifikatorjih za novo enoto določimo pripadnost tistemu razredu, kjer je verjetnost največja, ga lahko obravnavamo kot neverjetnostnega. Mere za vrednotenje uvrščanja so sprva razvili za uvrščanje v dva razreda. Mere za vrednotenje uvrščanja v več razredov pa so večinoma razvili kot posplošitve mer uvrščanja v dva razreda, nekatere pa so se razvile tudi neodvisno. Pri našem delu smo se osredotočili na mere za vrednotenje ne-verjetnostnih klasifikatorjev pri uvrščanju v več razredov.

#### 3.3.3.1 Mere vrednotenja uvrščanja za posamezen razred

1. **Napovedno točnost  $i$ -tega razreda** ( $PA_i$ , angl. class specific accuracy) izračunamo kot delež pravilno napovedanih enot  $i$ -tega razreda

$$PA_i = \frac{c_{ii}}{n_i}.$$

2. **Napovedno vrednost  $i$ -tega razreda** (PV, angl. Predictive Value) izračunamo kot delež pravilno napovedanih enot med vsemi enotami, napovedanimi v  $i$ -ti razred:

$$PV_i = \frac{c_{ii}}{n'_i}.$$

3. **F-mera  $i$ -tega razreda** (angl. per class F-score) je harmonična sredina napovedne točnosti in napovedne vrednosti  $i$ -tega razreda:

$$F\text{-mera}_i^\beta = (1 + \beta^2) \cdot \frac{PV_i \cdot PA_i}{(\beta^2 \cdot PV_i) + PA_i},$$

kjer je  $\beta$  koeficient, s katerim določamo relativno pomembnost napovedne vrednosti v primerjavi z napovedno točnostjo in je običajno enak 1:

$$F\text{-mera}_i^1 = 2 \cdot \frac{PV_i \cdot PA_i}{PV_i + PA_i}.$$

4. **Produkt napovedne vrednosti in napovedne točnosti  $i$ -tega razreda** [115] je geometrijska sredina napovedne vrednosti in napovedne točnosti za posamezen razred:

$$G\text{-mera}_i = \sqrt{PA_i \cdot PV_i}.$$

5. **Povprečna vrednost napovedne vrednosti in napovedne točnosti  $i$ -tega razreda** [115] je aritmetična sredina napovedne vrednosti in napovedne točnosti za posamezni razred:

$$A\text{-mera}_i = \frac{PA_i + PV_i}{2}.$$

6. **Jaccardov koeficient  $i$ -tega razreda** [115] [110] ( $J_i$ , angl. Jaccard's coefficient) izračunamo kot število pravilno napovedanih enot v  $i$ -tem razredu ulomljeno z vsoto  $i$ -te vrstice in  $i$ -tega stolpca v kontingenčni tabeli  $C$ , pri čemer diagonalni element štejemo le enkrat:

$$J_i = \frac{c_{ii}}{n_i + n'_i - c_{ii}}.$$

7. **Indeks uspešnosti uvrščanja v  $i$ -ti razred** [111] ( $CSI_i$ , angl. Individual Classification Success Index)

$$CSI_i = PA_i + PV_i - 1.$$

Vrednosti  $CSI_i$  so na intervalu  $[-1, 1]$ , pri čemer je  $-1$ , kadar sta obe napaki (napačno uvrščene enote iz  $i$ -tega razreda; v  $i$ -ti razred uvrščene enote, ki ne pripadajo  $i$ -temu razredu) maksimalni, in  $1$ , kadar sta obe napaki minimalni.

### 3.3.3.2 Mere za vrednotenje uvrščanja v več razredov, izračunane na podlagi kontingenčne tabele delovanja klasifikatorja

V tem razdelku so od točke 12. naprej predstavljene mere, ki so izpeljane iz informacijske entropije [158]. Pri teh merah je logaritem definiran tako, da je njegova vrednost v točki 0 enaka 0

$$\log_a 0 = 0.$$

1. **Napovedna točnost** [102] ( $PA$ , angl. Predictive Accuracy) je definirana kot delež pravilno napovedanih enot:

$$PA = \frac{\sum_{i=1}^K c_{ii}}{N}.$$

Vrednost  $PA$  je enaka tudi t. i. mikropovprečjem napovedne vrednosti, napovedne točnosti in F-mere.  $PA$  zavzame vrednost 0, ko so vse enoto uvrščene napačno, in vrednost 1, ko so vse enote uvrščene pravilno. Stopnja napake

$$E = 1 - PA$$

ima enake lastnosti kot  $PA$ , le da vrednost 0 predstavlja najboljši klasifikator, vrednost 1 pa najslabši.

2. **A-povprečje** [103] ( $A$ , angl. macro-average Aritmetic, v primeru dveh razredov tudi balanced accuracy -  $AUC_b$  [159]) je aritmetična sredina napovednih točnosti posameznih razredov

$$A = \frac{\sum_{i=1}^K PA_i}{K}.$$

Mera  $A$  zavzame vrednost 0, če so vse napovedne točnosti posameznih razredov enake 0, in vrednost 1, če so vse napovedne točnosti posameznih razredov enake 1.

3. **G-povprečje** [104] ( $G$ , angl. macro-average Geometric ali G-mean) je geometrijska sredina napovednih točnosti posameznih razredov

$$G = \left( \prod_{i=1}^K PA_i \right)^{\frac{1}{K}}.$$

Vrednost  $G$  je enaka 0, če je napovedna točnost le enega od razredov enaka 0, kar je v podatkih z močnim neravnotežjem in nekaterimi zelo majhnimi razredi lahko pogost pojav. To pomanjkljivost smo skušali odpraviti tako, da smo napovedne točnosti razredov, ki so bile enake 0, zamenjali z vrednostjo 0,01 in nato izračunali  $G$ -povprečje (oznaka:  $G-0.01$ ). Mera  $G$ -povprečje zavzame najvišjo vrednost 1, ko so napovedne točnosti vseh razredov enake 1.

4. **Kappa statistika** [98] ( $K$ ) meri, koliko je klasifikator boljši od naključnega klasifikatorja. Razlika med napovedno točnostjo klasifikatorja ( $PA$ ) in napovedno točnostjo naključnega klasifikatorja ( $RA$ , angl. Random Accuracy) je izražena kot delež razlike med napovedno točnostjo idealnega klasifikatorja (1) in naključnega klasifikatorja ( $RA$ ):

$$K = \frac{PA - RA}{1 - RA},$$

kjer je

$$RA = \frac{1}{N} \sum_{i=1}^K \frac{(n_i \cdot n'_i)}{N}.$$

V primeru, ko klasifikator uvrsti vse enote pravilno, doseže kappa največjo vrednost ( $K = 1$ ), najmanjšo vrednost ( $K = 0$ ) pa doseže v primeru, ko klasifikator uvrsti pravilno toliko enot, kot bi jih pravilno uvrstil naključni klasifikator.

5. **Makropovprečje napovedne vrednosti** [102] ( $MAP$ , angl. Macro-Averaged Precision) je aritmetična sredina napovednih vrednosti za posamezne razrede:

$$MAP = \frac{\sum_{i=1}^K PV_i}{K}.$$

V primeru, ko v  $i$ -ti razred ni uvrščena nobena enota, potem napovedna vrednost za  $i$ -ti razred ni definirana zaradi deljenja z nič. V našem algoritmu smo v takem primeru  $i$ -temu razredu priredili napovedno vrednost 0. Mera  $MAP$  zavzame najnižjo vrednost 0, ko v noben razred ni pravilno uvrščena nobena enota, in najvišjo vrednost 1, ko so vse enote uvrščene pravilno.

6. **F-povprečje** [160] ( $F$ , angl. mean F-measure) je aritmetična sredina  $F$ -mer za posamezne razrede

$$F = \frac{\sum_{i=1}^K F\text{-mera}_i}{K}.$$

$F$ -povprečje zavzame vrednosti med 0 in 1, pri čemer vrednost 1 pomeni popolno uvrščanje. V primeru, da v  $i$ -ti razred ni pravilno uvrščena nobena enota, vrednost  $F\text{-mera}_i$  ni definirana zaradi deljenja z 0. V našem algoritmu smo v takem primeru  $F$ -meri  $i$ -tega razreda ( $F\text{-mera}_i$ ) priredili vrednost 0.

7. **Korelacijski koeficient Matthew** ( $MCC$ , angl. Mathew's Correlation Coefficient,  $K$ -category correlation coefficient  $R_k$ ) je bil za uvrščanje v dva razreda predstavljen v [105] in posplošen za več razredov v [106].  $MCC$  je posplošitev Pearsonovega korelacijskega koeficienta na kategorične spremenljivke in ga lahko izračunamo iz kontingenčne tabele delovanja klasifikatorja:

$$MCC = \frac{N \cdot \sum_{i=1}^K c_{ii} - \sum_{i,j=1}^K c_{i \cdot} c_{\cdot j}}{\sqrt{N^2 - \sum_{i,j=1}^K c_{i \cdot} c_{\cdot j}} \sqrt{N^2 - \sum_{i,j=1}^K c_{i \cdot} c_{\cdot j}}},$$

kjer  $c_i$  predstavlja  $i$ -to vrstico  $c_{\cdot i}$  pa  $i$ -ti stolpec kontingenčne tabele.

Podobno kot Pearsonov korelacijski koeficient lahko MCC zavzame vrednosti na intervalu  $[-1, 1]$ . Vrednost 1 je dosežena, ko so vse enote uvrščene pravilno. Vrednost 0 je dosežena, ko klasifikator vse enote uvrsti v isti razred, ali ko so vse celice v kontingenčni matriki enake. Vrednost  $-1$  predstavlja ekstremen primer napačne klasifikacije, ko sta le dve simetrični enoti v kontingenčni tabeli različni od 0 ( $c_{ij} \neq 0$  in  $c_{ji} \neq 0$ ). V [90] so za uvrščanje v dva razreda pokazali, da ima MCC povezavo s  $\chi^2$  porazdelitvijo.

8. **Verjetnostna napovedna točnost** [107] (Pacc, angl. Probabilistic accuracy measure).  $P_{ij}$  predstavlja vsoto verjetnosti, ko klasifikator uvrsti enoto iz razreda  $i$  v razred  $j$  pri pogoju, da je dejanski razred  $i$ , in verjetnosti, ko klasifikator uvrsti enoto iz razreda  $i$  v razred  $j$  pri pogoju, da je napovedan razred  $j$ :

$$P_{ij} = \frac{2c_{ij}}{\sum_{l=1}^K (c_{il} + c_{lj})}; i, j = 1, 2, \dots, K.$$

Kot napačnost  $err$  in pravilnost  $corr$  označimo vsoti:

$$corr = \frac{1}{K} \sum_{i=1}^K P_{ii} \text{ in } err = \frac{1}{K} \sum_{i \neq j} P_{ij}.$$

Vrednost Pacc je potem definirana kot:

$$Pacc = \frac{1}{2} + \frac{corr - err}{2}.$$

Vrednosti Pacc ležijo na intervalu  $[0, 1]$ . Vrednost 1 pomeni, da je klasifikator uvrstil vse enote pravilno. Vrednost 0, pomeni, da klasifikator ni uvrstil pravilno nobene enote in je pri tem vse enote iz enega razreda uvrstil v isti "napačni" razred.

9. **Po razredih uravnotežena napovedna točnost** [108] (CBA, angl. Class Balance Accuracy). Za vsak razred  $i$  je izračunana  $CBA_i$  tako, da število pravilno uvrščenih enot  $i$ -tega razreda  $c_{ii}$  delimo s številom dejanskih enot v  $i$ -tem razredu  $n_i$  ali s številom v  $i$ -ti razred uvrščenih enot  $n'_i$ , odvisno od tega, katero od obeh števil je večje:

$$CBA_i = \frac{c_{ii}}{\max(n_i, n'_i)}.$$

Skupna CBA je povprečna vrednost  $CBA_i$ -jev:

$$CBA = \frac{\sum_{i=1}^K CBA_i}{K}.$$

Če so vse enote pravilno uvrščene v  $i$ -ti razred, potem je  $CBA_i = 1$ , če pa ni pravilno uvrščena nobena enota, potem je  $CBA_i = 0$ . Maksimum v imenovalcu poskrbi, da enkrat delimo z vsoto stolpcev drugič pa z vsoto vrstic v kontingenčni tabeli, s čimer preprečimo, da bi klasifikator, ki vse enote uvršča v isti razred, dobil visoko vrednost CBA.

10. **Jaccardov koeficient** [109, 110] (J, angl. Jaccard's coefficient, mean intersection over union). Skupni Jaccardov koeficient izračunamo kot povprečje Jaccardovih koeficientov posameznih razredov:

$$J = \frac{\sum_{i=1}^K J_i}{K}.$$

Mera  $J$  lahko zavzame vrednosti na intervalu  $[0, 1]$ , pri čemer je zgornja meja zavzeta, ko so vse enote uvrščene pravilno, spodnja pa, ko ni nobena enota uvrščena pravilno.

11. **Indeks uspešnosti uvrščanja** [111] (CSI, angl. Classification Success Index) izračunamo kot povprečni indeks uspešnosti uvrščanja posameznega razreda  $CSI_i$ :

$$CSI = \frac{\sum_i^K CSI_i}{K}.$$

V primeru, ko v  $i$ -ti razred ni uvrščena nobena enota, napovedna vrednost  $i$ -tega razreda ni definirana zaradi deljenja z nič. V našem algoritmu smo v takem primeru  $i$ -temu razredu priredili napovedno vrednost 0.

12. **Razvrstitvena entropija** [113] (CEN, angl. Confusion Entropy) je definirana s pomočjo verjetnosti in osnovnih idej informacijske teorije. Mera CEN upošteva kar največ informacije, ki se jo da pridobiti iz nediagonalnih elementov v kontingenčni tabeli. S  $P_{ij}^j$  označimo verjetnost, da klasifikator uvrsti element iz razreda  $i$  v razred  $j$  pri pogoju, da sta dejanski ali napovedani razred enaka razredu  $j$ :

$$P_{ij}^j = \frac{c_{ij}}{\sum_{l=1}^K (c_{jl} + c_{lj})},$$

in s  $P_{ij}^i$  verjetnost, da klasifikator uvrsti element iz razreda  $i$  v razred  $j$  pri pogoju, da sta dejanski ali napovedani razred enaka razredu  $i$ :

$$P_{ij}^i = \frac{c_{ij}}{\sum_{l=1}^K (c_{il} + c_{li})},$$

$$P_{ii}^i = 0.$$

Nato lahko definiramo

$$P_j = \frac{\sum_{l=1}^K (c_{jl} + c_{lj})}{2 \sum_{i,l=1}^K c_{il}}$$

in razvrstitveno entropijo  $j$ -tega razreda

$$CEN_j = - \sum_{l=1, l \neq j}^K \left( h_{2(K-1)}(P_{jl}^j) + h_{2(K-1)}(P_{lj}^j) \right),$$

kjer je

$$h_b(x) = \begin{cases} x \log_b x; & x \neq 0 \\ 0; & \text{sicer.} \end{cases}$$

Skupno razvrstitveno entropijo nato definiramo:

$$CEN = \sum_{j=1}^K CEN_j.$$

Za število razredov  $K > 2$  zavzame  $CEN$  vrednosti na intervalu  $[0, 1]$ . V nasprotju z večino drugih mer  $CEN = 0$  predstavlja klasifikator, ki uvršča brez napak in  $CEN = 1$  predstavlja klasifikator, ki napove vse enote napačno. V primeru, ko je število razredov  $K = 2$ , lahko mera  $CEN$  zavzame tudi vrednosti večje od 1. Zaradi lažjega primerjanja mere CEN z ostalimi merami so v rezultatih prikazane vrednosti  $1 - CEN$ .



13. **Relativna informacija klasifikatorja** [114] (RCI, angl. Relative Classifier Information). Negotovost problema je definirana kot

$$H_d = - \sum_{i=1}^K \frac{\sum_{j=1}^K c_{ij}}{N} \log \frac{\sum_{j=1}^K c_{ij}}{N}.$$

Skupna negotovost po opazovanju je definirana kot

$$H_O = \sum_{j=1}^K \frac{\sum_{i=1}^K c_{ij}}{N} H_{O_j},$$

kjer je

$$H_{O_j} = - \sum_{i=1}^K \frac{c_{ij}}{\sum_{i=1}^K c_{ij}} \log \frac{c_{ij}}{\sum_{i=1}^K c_{ij}}.$$

Negotovost, ki smo jo z modelom uvrščanja odstranili, potem izračunamo kot

$$H_c = H_d - H_O.$$

Kvocien

$$\text{RCI} = \frac{H_c}{H_d}$$

pa se imenuje relativna informacija klasifikatorja.

Mera RCI meri, kako dobro so razredi ločeni med seboj. Zavzame lahko vrednosti na intervalu  $[0, 1]$ , kjer 0 pomeni, da klasifikator med razredi ne razloči, 1 pa pomeni, da so vsi razredi med seboj popolnoma ločeni.

14. **Normalizirano vzajemno informacijo** [115, 116] (NMI, angl. Normalized Mutual Information) izračunamo kot

$$\text{NMI} = \frac{-2 \sum_{i,j=1}^K c_{ij} \log \left( \frac{c_{ij} N}{n_i n'_j} \right)}{\sum_{i=1}^K \left( n_i \log \left( \frac{n_i}{N} \right) + n'_i \log \left( \frac{n'_i}{N} \right) \right)}.$$

V primeru idealnega klasifikatorja je vrednost NMI enaka 1, v primeru klasifikatorja, ki uvrsti vse enote napačno, pa 0.

15. **Normalizirana prenesena informacija** [117, 118] (H, angl. normalized transmitted information). Prenesena informacija je definirana kot

$$h = \frac{1}{N} \sum_{i,j} c_{ij} \left( \log c_{ij} - \log n_i - \log n'_j + \log N \right).$$

Zavzame vrednosti večje od 0, zgornja meja je odvisna od števila razredov in števila enot v posameznem razredu. Zato vrednost  $h$  normaliziramo z vrednostjo prenesene informacije idealnega klasifikatorja  $h_{max}$ . Normalizirana prenesena informacija je potem enaka:

$$H = \frac{h}{h_{max}}$$

in zavzame vrednost 0 v primeru naključnega klasifikatorja ali v primeru, ko so vse enote uvrščene v isti razred, in vrednost 1, če so vse enote uvrščene pravilno.

### 3.3.4 pristopi za vrednotenje uvrščanja

Z ustreznim pristopom za vrednotenje uvrščanja želimo iz danih podatkov čim bolj natančno ovrednotiti uvrščanje novih – še neznanih enot. Če ocenimo določeno mero za vrednotenje uvrščanja na istih enotah, kot smo jih uporabili za gradnjo modela, je ocena zelo pristranska navzgor in zato nezanesljiva. Naslednji preprost pristop za ocenjevanje mere je, da podatke razdelimo na učno in testno množico, pri čemer na učni množici klasifikator zgradimo, na testni pa ga ovrednotimo. Pri tem pristopu se postavi vprašanje, kako učinkovito razdeliti podatke na učno in testno množico. Ta pristop je še posebej problematičen v primeru majhnega števila podatkov. Učna množica je potem, ko nekaj podatkov “žrtvujemo” za testno množico, majhna. Klasifikator, ki ga zgradimo, je tako že zaradi majhnosti učne množice manj učinkovit. Ocena napovedne točnosti pri tem pristopu je zato pristranska navzdol. Cilj pristopov za vrednotenje uvrščanja je torej zgraditi čim bolj učinkovit model uvrščanja tako, da za njegovo učenje uporabimo kar največ podatkov, ki jih imamo na razpolago, in kljub temu ovrednotimo uvrščanje na enotah, ki niso odvisne od enot, uporabljenih za izgradnjo modela.

V naših analizah smo za vrednotenje uvrščanja uporabili večkrat ponovljeno stratificirano razdelitev na učno in testno množico in večkrat ponovljeno prečno preverjanje s pregibanjem.

#### 3.3.4.1 Večkrat ponovljena stratificirana razdelitev na učno in testno množico (angl. stratified split-sampling)

Pri stratificirani razdelitvi na učno in testno množico podatke razdelimo na dve disjunktni množici. Testno množico sestavljajo enote, ki jih iz vsakega razreda izberemo naključno na takšen način, da so razmerja velikosti razredov testne množice enaka razmerjem v začetnih podatkih. Enote, ki jih nismo izbrali v testno množico, pa sestavljajo učno množico. Na učni množici zgradimo klasifikator in mero za vrednotenje uvrščanja ocenimo na testni množici. Če želimo zmanjšati vpliv razdelitve podatkov na učno in testno množico, in s tem zmanjšati variabilnost rezultatov vrednotenja uvrščanja, lahko postopek večkrat ponovimo [87], dobljene ocene mer za vrednotenje uvrščanja pa na koncu povprečimo.

V naših izračunih smo postopek stratificiranega vzorčenja ponovili 500 krat, podatke pa smo na vsakem koraku razdelili v učno in testno množico približno v razmerju 2 : 1.

#### 3.3.4.2 Prečno preverjanje s pregibanjem

Pri prečnem preverjanju z  $L$  pregibi (CV, angl. *L-fold Cross-Validation*) podatke naključno razdelimo na  $L$  disjunktih podmnožic. Postopek nato poteka v  $L$  korakih. V  $l$ -tem koraku zgradimo model uvrščanja na učni množici, ki je sestavljena iz vseh podatkov, razen tistih iz  $l$ -te podmnožice. S pomočjo naučenega klasifikatorja nato napovemo enote iz  $l$ -te podmnožice. Na takšen način enote, ki jih napovedujemo, niso hkrati vključene v gradnjo klasifikatorja. Po  $L$  korakih dobimo napovedi za vse enote, ki jih lahko potem uporabimo za izračun mere za vrednotenje uvrščanja. Če želimo zmanjšati vpliv prvotne razdelitve podatkov na  $L$  disjunktih podmnožic in s tem zmanjšati variabilnost rezultatov vrednotenja uvrščanja, lahko postopek večkrat ponovimo [87, 161].

V doktorskem delu smo uporabljali 100 krat ponovljeno prečno preverjanje z 10 pregihi.

### 3.3.5 Metode ocenjevanja skladnosti mer za vrednotenje uvrščanja

Metode za ocenjevanje skladnosti so v doktorskem delu uporabljene z namenom, da se ovrednoti skladnost delovanja mer za vrednotenje uvrščanja, ki so bile predstavljene v razdelku 3.3.3.

#### 3.3.5.1 Kendallov korelacijski koeficient

Kendallov koeficient korelacije rangov  $\tau$  [162] meri skladnost rangov dveh spremenljivk  $X$  in  $Y$ . Naj bodo  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  opazovane vrednosti spremenljivk  $X$  in  $Y$ . Par  $(x_i, y_i)$  in  $(x_j, y_j)$ , kjer je  $i \neq j$ , je *skladen* (angl. concordant), če je  $(x_i > x_j) \wedge (y_i > y_j)$  ali  $(x_i < x_j) \wedge (y_i < y_j)$ . Par je *neskladen* (angl. discordant), če je  $(x_i > x_j) \wedge (y_i < y_j)$  ali  $(x_i < x_j) \wedge (y_i > y_j)$ . V primeru, ko je  $x_i = x_j$  ali  $y_i = y_j$  par ni niti skladen niti neskladen. Kendallov korelacijski koeficient izračunamo kot

$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}},$$

kjer je  $n_c$  število skladnih parov in  $n_d$  število neskladnih parov.

V primeru vezanih rangov ga izračunamo kot

$$\tau = \frac{n_c - n_d}{\left(\frac{n(n-1)}{2} - t_x\right) \left(\frac{n(n-1)}{2} - t_y\right)},$$

kjer je  $t_x$  število vezanih rangov pri opazovanjih spremenljivke  $X$ ,  $t_y$  pa število vezanih rangov pri opazovanjih spremenljivke  $Y$ .

Kendallov  $\tau$  lahko zavzame vrednosti med  $-1$  in  $1$ , kjer vrednost  $1$  predstavlja popolno skladnost rangov vrednost  $-1$  pa popolno neskladje rangov.

#### 3.3.5.2 Spearmanov korelacijski koeficient

Spearmanov koeficient korelacije rangov  $\rho$  med dvema spremenljivkama izračunamo kot

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

kjer  $d_i$  označuje razlike med rangi istoležnih členov obeh spremenljivk,  $n$  pa je število enot. Spearmanov korelacijski koeficient zavzame vrednosti med  $-1$  in  $1$ , kjer vrednost  $-1$  pomeni, da med spremenljivkama obstaja popolna (monotona) negativna korelacija, vrednost  $0$  pove, da med spremenljivkama ni monotone korelacije, vrednost  $1$  pa, da je med spremenljivkama popolna (monotona) pozitivna korelacija.

### 3.3.5.3 Kendallov koeficient konkordance

Kendallov koeficient konkordance ( $W$ ) meri skladnost rangov  $p$  spremenljivk (oz. usklajenost  $p$  ocenjevalcev). Izračunamo ga lahko s pomočjo Spearmanovih korelacijskih koeficientov [163]:

$$W = \frac{(p-1)\bar{r} + 1}{p},$$

kjer je  $p$  število spremenljivk (ocenjevalcev),  $\bar{r}$  pa je aritmetična sredina Spearmanovih korelacijskih koeficientov med vsemi pari spremenljivk. Kendallov  $W$  zavzame vrednosti med 0 in 1, kjer 0 predstavlja popolno neskladje, 1 pa popolno usklajenost spremenljivk.

### 3.3.5.4 Prikaz konkordance z rangi na vzporednih oseh

Usklajenost oz. konkordanco  $p$  spremenljivk (ocenjevalcev) na  $n$  enotah lahko prikažemo tako, da rangiramo vrednosti pri vsaki spremenljivki, nato prikažemo izračunane range na  $p$  vzporednih oseh, ki predstavljajo spremenljivke. Točke, ki predstavljajo range iste enote, povežemo. Tako dobimo  $n$  črt na  $k$  vzporednih oseh, pri čemer več presečišč predstavlja manjšo konkordanco, manj presečišč pa večjo [164, 165].

### 3.3.5.5 Konkordančni mehurčni diagram

Za prikaz konkordance  $p$  spremenljivk (ocenjevalcev) na  $n$  enotah s konkordančnim mehurčnim diagramom [165] rangiramo vrednosti pri vsaki spremenljivki. Za vsako enoto nato izračunamo povprečni rang in frekvence rangov. Na abscisno os nanašamo range, na ordinatno os pa povprečne range. Za vsako enoto narišemo pri vrednosti ordinatne osi, ki je enaka povprečnemu rangi enote, toliko krogov, kot je različnih rangov. Ploščine narisanih krogov so sorazmerne frekvenci rangov. Če vsi krogi ležijo na glavni diagonali, imamo popolno konkordanco. Bolj kot so krogi razpršeni, manjša je konkordanca.

## 3.4 Opis generiranja podatkov

Za testiranje delovanja novih statističnih metod in primerjanje učinkovitosti med metodami se je pojavila potreba po simulacijah oz. generiranju “umetnih podatkov”. V nasprotju z realnimi podatki lahko simuliranih podatkov generiramo poljubno mnogo, ob tem pa poznamo njihove lastnosti, zaradi česar lahko tako manj pristransko in natančneje ocenimo delovanje metod. Za generiranje podatkov pogosto uporabljamo znane porazdelitvene funkcije, s katerimi le v redkih primerih dobimo podatke, primerljive realnim. Pri visoko razsežnih podatkih se ta problem še stopnjuje, saj je zelo težko pravilno simulirati zapleteno korelacijsko strukturo med spremenljivkami.

V nadaljevanju so predstavljeni štirje načini generiranja NIRS podatkov. Pri prvem in drugem načinu so iz realnih podatkov ocenjene razlike med razredi, pri tretjem načinu je iz realnih podatkov ocenjena tudi korelacijska struktura. Četrty način generiranja podatkov se od prvih treh bistveno razlikuje, saj smo poskusili podatke generirati na podlagi teoretičnih, kemijskih in fizikalnih, zakonitosti.

Pri generiranju NIRS podatkov smo se osredotočili na tri vrste polimerov polietilen (PE,  $n_{PE} = 26$ ), polipropilen (PP,  $n_{PE} = 27$ ) in polistiren (PS,  $n_{PS} = 59$ ), ki so v naši zbirki dejanskih podatkov (Tabela 3.1.1) in tudi sicer v zgodovinskih zbirkah polimerov med pogostejšimi. Te tri vrste polimerov imajo tudi dovolj preprosto kemijsko strukturo (razdelek 2.1.4), da smo zanje lahko pridobili zadostno količino sorazmerno natančnih informacij o absorpcijah v NIR območju, ki smo jih potrebovali v eni od metod generiranja podatkov. PE in PP sta si po kemijski strukturi bolj podobna kot PE in PS ali PP in PS. Ker je kemijska struktura snovi v tesni povezavi z absorpcijo svetlobe v NIR območju, lahko to (ne)podobnost opazimo tudi na slikah spektrov (Slika 2.1.1).

Pri generiranju neodvisnih podatkov, pridobljenih s pomočjo univariatne normalne porazdelitve  $N(\mu, \sigma)$  (razdelek 3.4.1), in pri generiranju podatkov s pomočjo multivariatne normalne porazdelitve  $MVN(M, \Sigma)$ , kjer smo povezanost spremenljivk simulirali s pomočjo bločne kovariančne matrike (razdelek 3.4.2), smo se srečali z vprašanjem, kako definirati razlike med razredi, da bodo le-te primerljive z realnimi podatki. Na primer, ko smo za parametra  $\mu$  in  $M$  izbrali preproste kombinacije z vrednostma 0 in 1, so se generirani podatki izkazali za popolnoma neprimerljive z realnimi podatki in zato neuporabne (ni prikazano v doktorskem delu). Parametra  $\mu$  in  $M$  smo nato določili tako, da so bili rezultati uvrščanja uravnoveženih podatkov v dva razreda pri simuliranih podatkih približno enaki kot pri realnih podatkih. Vendar smo ob uporabi drugačne metode uvrščanja spet naleteli na isti problem. Zato smo bili primorani za vsako od obravnavanih metod uvrščanja določiti drugačne vrednosti parametrov  $\mu$  in  $M$ , ki so prikazane v Tabeli 4.2.1 v poglavju Rezultati.

Tudi pri drugih dveh načinih generiranja podatkov, se opiramo na nekatere rezultate, zato je v tem razdelku metodologija generiranja podatkov opisana bolj jedrnato, podrobnosti metod, ki se nanašajo na rezultate, pa so opisane v poglavju Rezultati v razdelku 4.2.

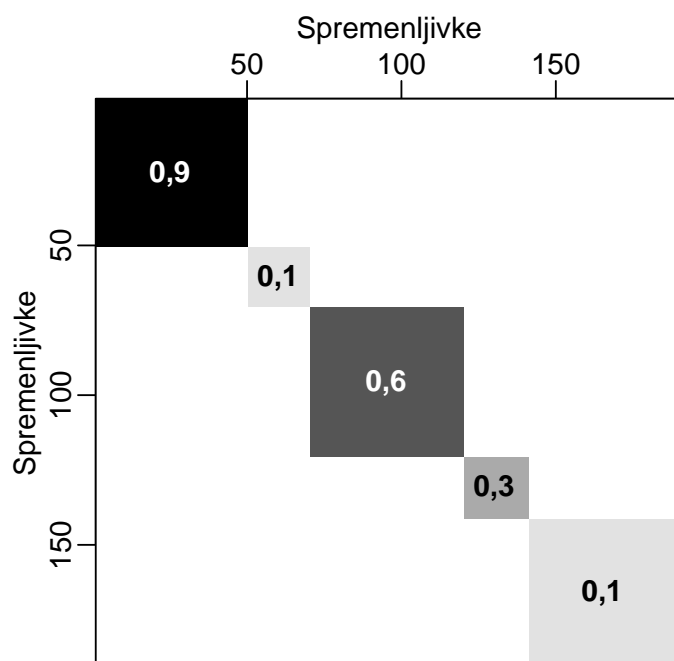
### 3.4.1 Generiranje neodvisnih podatkov (IND)

Spremenljivke smo generirali neodvisno iz univariatne normalne porazdelitve  $N(\mu, \sigma)$ . Parameter  $\sigma$  je bil za vse spremenljivke v vseh razredih enak 1, medtem ko smo razlike med

razredi generirali s pomočjo razlik v parametru  $\mu$ . Vse spremenljivke iz skupine PE in večina spremenljivk iz skupin PP in PS je bilo generiranih s parametrom  $\mu = 0$ . V skupini PP je bilo 10 spremenljivk, generiranih s parametrom  $\mu = \mu_{PP} \neq 0$  in v skupini PS je bilo 30 spremenljivk, generiranih s parametrom  $\mu = \mu_{PS} \neq 0$ . Vrednosti  $\mu_{PP}$  in  $\mu_{PS}$  sta bili pri vsaki metodi uvrščanja določeni tako, da so se vrednosti G-povprečja pri uvrščanju v dva uravnotežena razreda ujemale z realnimi podatki (Tabela 4.2.1).

### 3.4.2 Generiranje koreliranih spremenljivk z bločno kovariančno matriko (MVNblock)

Spremenljivke smo generirali iz multivariatne normalne porazdelitve  $MVN(M, \Sigma)$ , kjer je  $M$  vektor povprečij in  $\Sigma$  kovariančna matrika. Kovariančna matrika je bila sestavljena iz petih blokov, kar je predstavljalo pet skupin spremenljivk, ki so bile med seboj pozitivno korelirane, medtem ko so bile spremenljivke iz različnih skupin neodvisne. Struktura kovariančne matrike je bila za vse skupine (PE, PP in PS) enaka in je prikazana na Sliki 3.4.1.



Slika 3.4.1: Kovariančna matrika, uporabljena pri generiranju koreliranih spremenljivk z bločno kovariančno matriko.

Pri generiranju podatkov iz skupine PE je bil vektor povprečij  $M = 0$ . Pri skupini PP smo naključno izbrali 10 komponent, pri skupini PS pa 30 komponent vektorja povprečij  $M$ , ki so imele vrednosti različne od 0. Vrednosti teh različno izraženih komponent smo pri vsaki skupini in vsaki metodi uvrščanja določili tako, da so se vrednosti G-povprečja pri uvrščanju v dva uravnotežena razreda ujemale z realnimi podatki (Tabela 4.2.1).

### 3.4.3 Generiranje podatkov iz multivariatne normalne porazdelitve s parametri, ocenjenimi iz realnih podatkov (MVNorig)

Podatke smo generirali iz multivariatne normalne porazdelitve  $MVN(M, \Sigma)$ , kjer  $M$  predstavlja vektor povprečij,  $\Sigma$  pa kovariančno matriko. Parametra  $M$  in  $\Sigma$  sta bila ocenjena na podlagi vzorčnih vrednosti za vsako skupino (PE, PP in PS) posebej iz dejanskih podatkov polimerov.

### 3.4.4 Generiranje podatkov na podlagi teoretičnih absorpcij (ABS)

V tem poglavju predstavljamo nov pristop za generiranje spektralnih podatkov, ki predstavljajo odboj svetlobe v NIR območju. Podatke smo generirali po funkciji

$$odboj(x_i^G) = t_i(x) - C_i \cdot (TA^G(x) + RA_i(x)) + \epsilon(x), \quad (3.3)$$

kjer je  $i$  enota,  $G$  skupina,  $x$  meritev (valovna dolžina),  $TA^G(x)$  teoretična absorpcija skupine,  $RA_i(x)$  so dodatna absorpcijska mesta,  $t_i(x)$  je trendna funkcija,  $C_i$  je konstanta in  $\epsilon(x)$  napaka posamezne meritve. V funkciji  $odboj(x_i^G)$  od trendne funkcije  $t_i(x)$  odštejemo s konstanto  $C_i$  pomnožen izraz  $(TA^G(x) + RA_i(x))$ , ki predstavlja absorpcijo. Na tak način smo iz podatkov o absorpciji simulirali spektre odboja.

V realnih podatkih spektri odboja nikoli ne zavzamejo negativnih vrednosti, zato smo v primeru, ko smo s predlagano funkcijo dobili negativne vrednosti, spekter popravili:

$$odbojAdj(x_i) = odboj(x_i^G) - \min(x_i) \cdot \left(1 + \frac{1}{\max(x_i)}\right),$$

pri čemer  $\min$  predstavlja najmanjšo,  $\max$  pa največjo vrednost spektra  $odboj(x_i^G)$ .

#### 3.4.4.1 Trend $t_i(x)$

Po natančnem opazovanju trenda realnih spektrov NIR območja (razdelek 2.1) smo trend modelirali kot eksponentno funkcijo  $t(x) = \exp(\alpha x + \beta)$  (Slika 4.2.5). Porazdelitev vektorja parametrov  $(\alpha, \beta)$  smo ocenili iz zbirke 534 spektrov odboja v NIR območju, ki so bili posneti na različnih polimerih, kot je opisano v razdelku 3.1. Za vsak spekter  $s \in \{1, 2, \dots, 534\}$  smo po metodi najmanjših kvadratov ocenili funkcijo eksponentnega trenda:  $t_s(x) = \exp(a_s x + b_s)$ . Nato smo ocenili povprečni vektor  $(\bar{a}, \bar{b})$  in variančno-kovariančno matriko  $\Sigma_{(a,b)}$ .

Za vsako na novo generirano enoto  $i$  smo trendno funkcijo  $t_i(x)$  v Enačbi 3.3 določili kot eksponentno funkcijo  $t_i(x) = \exp(a_i x + b_i)$ , kjer smo  $a_i$  in  $b_i$  generirali kot naključni števili iz multivariatne normalne porazdelitve  $MVN(M_{(\alpha,\beta)}, \Sigma_{(\alpha,\beta)})$ , kjer sta bila parametra  $M_{(\alpha,\beta)} = (\bar{a}, \bar{b})$  in  $\Sigma_{(\alpha,\beta)} = \frac{1}{2}\Sigma_{(a,b)}$ .

#### 3.4.4.2 Konstanta $C_i$

Konstanto  $C_i$  smo za vsako na novo generirano enoto  $i$  izračunali kot povprečno vrednost trendne funkcije na danem območju  $C_i = \text{mean}(t_i(x))$ . S tem smo dosegli, da imajo spektri

z na splošno višjim odbojem bolj izrazite absorpcije, kar je vidno tudi v realnih podatkih (Slika 2.1.1).

### 3.4.4.3 Teoretična absorpcija polimera $TA^G(x)$

Teoretično absorpcijo skupine smo določili za vsako vrsto plastik posebej, pri tem smo se opirali na informacije o absorpcijah posameznih funkcionalnih skupin v NIR območju, ki jih je mogoče najti v literaturi. V ta namen smo preiskali kemijsko sestavo polimerov in iz literature izpisali absorpcijska mesta (valovne dolžine) in moči absorpcijskih vrhov posameznih funkcionalnih skupin, ki sestavljajo polimere PE, PP in PS (razdelek 2.1.4). Kot glavni vir podatkov o absorpcijah funkcionalnih skupin smo uporabljali [57], manjkajoče oz. nepopolne informacija pa dopolnili s pomočjo [63–65, 67]. Moč absorpcij je težko natančno določiti, zato v literaturi pogosto ni podana, ob tem oznake za moč absorpcij med viri niso usklajene. V naših podatkih smo različne moči označili kot: zelo šibka (vw), šibka (w), srednje šibka (mw), srednja (m), srednje močna (ms) in močna (s), ki smo jih v istem zaporedju utežili z 0,05; 0,20; 0,35; 0,50; 0,75 in 1. Za vsako navedeno absorpcijsko mesto smo izračunali absorpcijsko krivuljo po Gaussovi funkciji (razdelek 2.1)

$$f(x) = A \cdot \exp\left(\frac{-(x - m)^2}{2 \cdot s^2}\right), \quad (3.4)$$

kjer  $A$  predstavlja moč absorpcije,  $m$  mesto največje absorpcije,  $s$  širino krivulje, spremenljivka  $x$  pa valovne dolžine NIR območja. Ker širina v literaturi ni podana, smo za vsa absorpcijska mesta izbrali enotno vrednost parametra  $s$ , to je 12. Skupno absorpcijsko krivuljo funkcionalne skupine  $FA(x)$  smo izračunali kot vsoto absorpcijskih krivulj ( $f(x)$ ) posameznih absorpcijskih mest (Slika 4.2.4).

Skupno teoretično absorpcijo posameznega polimera  $TA^G(x)$  iz Enačbe 3.3 smo izračunali za vsak polimer posebej kot glajeno uteženo vsoto krivulj, ki predstavljajo absorpcijo tistih funkcionalnih skupin, ki so sestavni deli polimera.

$$TA^G(x) = \text{smooth} \left( \sum_{j=1}^f w_j FA_j(x) \right), \quad (3.5)$$

kjer je smooth funkcija glajenja,  $w_j$  so uteži,  $FA_j$  pa skupna absorpcija  $j$ -te funkcionalne skupine in  $f$  število funkcionalnih skupin, ki sestavljajo polimer  $G$ . Za funkcijo glajenja smo uporabili Savitzky-Golay filter s polinomom stopnje 3 in oknom velikosti 7, pri čemer smo v programu **R** uporabili funkcijo *sgolayfilt* iz paketa *signal* [142].

Uteži  $w_j$  smo določili v povezavi z deležem, ki ga  $j$ -ta funkcionalna skupina predstavlja v molekuli polimera, saj se moč absorpcije funkcionalne skupine povečuje sorazmerno s številom, kolikokrat se funkcionalna skupina v molekuli pojavi [61]. Deležev metilne, metilenske in arilne funkcionalne skupine v molekulah PE, PP in PS ne moremo natančno določiti, saj nimamo informacije o dolžini verige in razvejanosti molekul, ki se lahko razlikujeta med različnimi vzorci istega polimera. Teoretično absorpcijo PE smo izračunali z utežjo 0,8 za metilensko in 0,2 za metilno skupino. Pri PP smo določili vrednost uteži pri metilenu 0,4, pri metilu pa 0,6. Pri PS je bila utež pri metilenski skupini 0,3, pri metilni 0,1 in pri



arilni 0,6. Na Sliki 4.2.6 so prikazane teoretične absorpcije posameznih funkcionalnih skupin pomnožene z ustreznimi utežmi in skupne absorpcije skupin PE, PP in PS.

#### 3.4.4.4 Dodatna absorpcijska mesta $RA_i(x)$

Ker vsaka enota vsebuje tudi svoje specifične (dodane so ji snovi – plastifikatorji, je bolj ali manj degradirala s časom, ima svojo obliko, površino, barvo ipd.), smo kot simulacijo teh specifik za vsako enoto naključno izbrali dodatni dve mesti  $m$ , na katerih smo absorpcijo izračunali po Enačbi 3.4. Parameter  $A$  iz Enačbe 3.4, ki predstavlja moč absorpcije, smo generirali kot naključno število iz enakomerne porazdelitve na intervalu od  $-0,3$  do  $0,2$ , kar pomeni, da na izbranih mestih  $m$  absorpcija ni bila nujno povečana, ampak je bila lahko tudi zmanjšana (pozitivne/negativne vrednosti). Parameter  $s$  smo generirali kot naključno število iz enakomerne porazdelitve na intervalu od  $4$  do  $200$ . Interval je širok z razlogom, da se vpliv specifik posamezne enote lahko opazi zelo lokalno (majhne vrednosti parametra  $s$ ) ali na večjem območju (visoke vrednosti parametra  $s$ ).

#### 3.4.4.5 Napaka posamezne meritve $\epsilon(x)$

Napaka posamezne meritve  $\epsilon(x)$  predstavlja napako merilnega aparata in druge slučajne vplive. Generirali smo jo kot naključno število iz normalne porazdelitve  $N(0; 0,2)$ .

### 3.5 Nastavitve simulacij

V tem razdelku so opisane nastavitve simulacij uvrščanja v dva, tri in 45 razredov. S simulacijami smo se želeli čim bolj približati realnemu problemu uvrščanja plastik, saj smo želeli rezultate uvrščanja umetno generiranih podatkov primerjati z rezultati uvrščanja realnih podatkov. Število generiranih enot v posamezni skupini je bilo zato takšno kot v realnih podatkih.

#### 3.5.1 Uvrščanje v dva in tri razrede

Podatke smo umetno generirali na štiri načine, opisane v razdelku 3.4. Ob generiranih podatkih smo uporabili tudi realne podatke. Tako realni podatki kot generirani podatki so bili sestavljeni iz skupin PE, PP in PS z velikostjo  $n_{PE} = 26$ ,  $n_{PP} = 27$  in  $n_{PS} = 59$ , kjer so bile med skupinama PE in PP manjše razlike kot med skupinama PE in PS ali PP in PS. Kemijska zgradba polimerov, iz katere lahko razberemo velikost razlik med razredi, je opisana v razdelku 2.1.4.

Pri gradnji modelov uvrščanja smo uporabili:

- 3 metode predobdelave: brez predobdelave, SNV in prvi odvod izračunan po metodi Savitzky-Golay;
- 4 metode zmanjšanja dimenzije podatkov: brez zmanjšanja dimenzije, izbor 50 spremenljivk z največjo varianco, izbor 50 spremenljivk z največjo F-statistiko, PCA;

- 2 metodi za zmanjšanje vpliva neravnotežja: brez zmanjšanja vpliva neravnotežja, MDS;
- 3 metode za uvrščanje podatkov: LDA, CART, L-SVM (metoda SVM, pri kateri smo uporabili linearno jedrno funkcijo in prilagojen prag za uvrščanje glede na velikost razredov).

### 3.5.1.1 Dva razreda

Modeli uvrščanja so bili zgrajeni za uvrščanje v dva razreda brez razlik (PS-PS), z majhnimi razlikami med razredoma (PE-PP) in z velikimi razlikami med razredoma (PE-PS). Pri uvrščanju v dva enaka razreda smo enote v skupini PS naključno razdelili v dve skupini s 30 in 29 enotami. V simulacijah smo ob razlikah med razredoma spreminjali tudi stopnjo neravnotežja, ki je bila določena z deležem enot prvega razreda  $k_1 \in \{0,1; 0,2; \dots; 0,9\}$ .

Pri generiranju podatkov po metodi MVNorig (razdelek 3.4.3) smo ocenili parametre multivariatne normalne porazdelitve iz realnih podatkov. Da bi izključili odvisnost podatkov MVNorig od realnih podatkov, vključenih v modele uvrščanja, smo za primer brez predobdelave in brez zmanjšanja dimenzije podatkov v vsaki od skupin PS, PP in PE naključno izbrali 10 enot, ki smo jih uporabili za oceno parametrov multivariatne porazdelitve v postopku generiranja MVNorig podatkov in jih kasneje nismo vključili v gradnjo modelov uvrščanja. Velikosti skupin, ki smo jih uporabili v simulacijah, so bile v tem primeru  $n_{PS} = 49$ ,  $n_{PP} = 17$  in  $n_{PE} = 16$ .

Pri metodi uvrščanja SVM smo s simulacijami preverili tudi vpliv jedrne funkcije in vpliv prilagoditve na velikost razredov. V ta namen smo uporabili dve jedrni funkciji: linearno in radialno. S prilagoditvijo metode SVM glede na velikost razredov smo želeli zmanjšati vpliv neravnotežja na rezultate uvrščanja, zato smo ob gradnji klasifikatorja s pomočjo privzetih nastavitev v funkciji *svm* iz paketa *e1071* uporabili še dve drugi možnosti:

1. v funkcijo *svm* vgrajeno možnost uteževanja razredov *class.weights*, kjer smo razred z deležem enot  $k_1$  utežili z utežjo  $w = 1 - k_1$ ;
2. prilagojen prag za uvrščanje (razdelek 3.3.2.3).

Pri vseh uporabljenih metodah in kombinacijah metod za gradnjo klasifikatorja smo za vsak delež  $k_1 \in \{0,1; 0,2; \dots; 0,9\}$  in za vsako od 1000 ponovitev (100 ponovitev ob uporabi metode MDS) izvedli naslednje korake:

1. iz prvega razreda smo naključno izbrali  $n_1$  in iz drugega  $n_2$  enot tako, da je veljalo  $\text{round}\left(\frac{n_1}{n_1+n_2}\right) = k_1$ , kjer je *round* funkcija, ki zaokroži število na najbližje celo število;
2. izbrane enote smo uporabili za izgradnjo modela uvrščanja, ki smo ga ovrednotili s pomočjo prečnega preverjanja z 10 pregibi;
3. izračunali smo napovedni točnosti razredov in vrednost G-povprečja.

Na koncu smo iz vseh 1000 ponovitev izračunali povprečne napovedne točnosti razredov in povprečno vrednost G-povprečja.

### 3.5.1.2 Trije razredi

Tudi pri simulacijah uvrščanja v tri razrede smo spreminjali razlike med razredi in neravnotežje. Modeli uvrščanja so bili zgrajeni na naslednjih kombinacijah razredov:

1. PS-PS-PS: vsi trije razredi so bili enaki, pri tem smo enote v skupini PS ( $n_{PS} = 59$ ) naključno razdelili v tri skupine velikosti  $n_{PS_1} = n_{PS_2} = 20$  in  $n_{PS_3} = 19$ ;
2. PP-PP-PE: prva dva razreda sta bila enaka, tretji razred pa se je z majhnimi razlikami razlikoval od prvih dveh, pri tem smo enote v skupini PP ( $n_{PP} = 27$ ) naključno razdelili v dve skupini velikosti  $n_{PP_1} = 14$  in  $n_{PP_2} = 13$ ;
3. PS-PS-PE: prva dva razreda sta bila enaka, tretji razred pa se je z velikimi razlikami razlikoval od prvih dveh, pri tem smo enote v skupini PS ( $n_{PS} = 59$ ) naključno razdelili v dve skupini velikosti  $n_{PS_1} = 30$  in  $n_{PS_2} = 29$ ;
4. PS-PP-PE: vsi trije razredi so bili različni, razlika med PS in PP je bila večja kot med PP in PE.

Pri vsaki kombinaciji razredov smo modele uvrščanja zgradili na razredih z velikostmi:

1.  $n_1 = 10, n_2 = 10, n_3 = 10$ ;
2.  $n_1 = 10, n_2 = 10, n_3 = 5$ ;
3.  $n_1 = 10, n_2 = 5, n_3 = 5$ ;
4.  $n_1 = 15, n_2 = 10, n_3 = 5$ .

Ker smo želeli preveriti tudi vpliv števila enot na uvrščanje, smo pri uvrščanju v tri enake razrede (PS-PS-PS) število enot v razredih povečali, tako da smo modele uvrščanja zgradili še za naslednje velikosti razredov:

1.  $n_1 = 20, n_2 = 20, n_3 = 10$ ;
2.  $n_1 = 20, n_2 = 10, n_3 = 10$ ;
3.  $n_1 = 20, n_2 = 10, n_3 = 5$ .

Za vsako kombinacijo razredov, za vsako kombinacijo velikosti razredov in za vsako od 1000 ponovitev so bili izvedeni naslednji koraki:

1. iz prve skupine smo naključno izbrali  $n_1$ , iz druge  $n_2$  in iz tretje  $n_3$  enot;
2. izbrane enote smo uporabili za izgradnjo modela uvrščanja, ki smo ga ovrednotili s pomočjo prečnega preverjanja z 10 pregibi;
3. izračunali smo napovedni točnosti razredov in vrednost G-povprečja.

Na koncu smo iz vseh 1000 ponovitev izračunali povprečne napovedne točnosti po razredih in povprečno vrednost G-povprečja.

### 3.5.2 Uvrščanje v 45 razredov

Pri simulacijah uvrščanja v 45 razredov smo uporabili generirane podatke po metodi MVNorig (razdelek 3.4.3). Razlike med razredi so bile določene s parametri MVN porazdelitve in smo jih za vsak razred ocenili iz realnih podatkov 45 vrst polimerov, ki so

opisani v razdelku 3.1. Število generiranih enot v posameznem razredu je bilo kot v realnih podatkih.

Pri gradnji modelov uvrščanja smo uporabili

- 3 metode predobdelave: brez predobdelave, SNV in prvi odvod izračunan po metodi Savitzky-Golay;
- 4 metode zmanjšanja dimenzije podatkov: brez zmanjšanja dimenzije, izbor 50 spremenljivk z največjo varianco, izbor 50 spremenljivk z največjo F-statistiko in PCA;
- 3 metode za uvrščanje podatkov: LDA, CART in L-SVM (metoda SVM, pri kateri smo uporabili linearno jedrno funkcijo in prilagojen prag za uvrščanje glede na velikost razredov).

Modele smo ovrednotili s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi in merama A-povprečje in G-povprečje.

## Poglavje 4

# Rezultati

Poglavje Rezultati je razdeljeno na tri dele. V prvem delu 4.1 so predstavljeni rezultati uvrščanja realnih podatkov polimerov. V drugem delu 4.2 so predstavljeni umetno generirani spektri. V tretjem delu 4.3 pa so predstavljeni rezultati simulacij uvrščanja realnih in umetno generiranih podatkov.

### 4.1 Uvrščanje dejanskih podatkov polimerov

V tem razdelku so prikazani rezultati uvrščanja dejanskih podatkov 45 vrst polimerov, pri čemer so bile uporabljene različne metode za predobdelavo, uvrščanje, zmanjšanje dimenzije podatkov, zmanjšanje vpliva neravnotežja in vrednotenje uvrščanja, kot je opisano v razdelku 3.1.

Razdelek je razdeljen na pet delov, kjer smo v prvem delu (razdelek 4.1.1) predstavili ocene mer vrednotenja uvrščanja za modele pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi in v drugem delu (razdelek 4.1.2) ocene mer vrednotenja uvrščanja za modele, pridobljene s 500 krat ponovljeno stratificirano razdelitvijo na testno in učno množico. Pri gradnji modelov uvrščanja smo v obeh delih uporabili eno od metod predobdelave (brez, kvantilna normalizacija, 1. odvod, SNV), eno od metod uvrščanja (LDA, CART, R-SVM, L-SVM, 1-NN, 3-NN, 5-NN) in eno od metod zmanjšanja dimenzije podatkov (brez, izbor spremenljivk na podlagi največje variance, izbor spremenljivk na podlagi največje F-statistike in zmanjšanje dimenzije z metodo PCA). Prva dva dela imata enako strukturo: na začetku obeh delov so predstavljene vrednosti aritmetične sredine napovednih točnosti posameznih razredov (A-povprečje) za vse obravnavane klasifikacijske modele. V nadaljevanju obeh poglavij so prikazane vrednosti različnih mer za vrednotenje uvrščanja (razdelek 3.3.3), pri čemer so rezultati modelov uvrščanja, ki so bili zgrajenih z metodo najbližjih sosedov z več kot enim sosedom, prikazani v dodatku. Izkazalo se je, da mere za vrednotenje uvrščanja delujejo skladno, zato smo se pri odločanju o najboljšem modelu osredotočili na mero A-povprečje.

V tretjem delu (razdelek 4.1.3) je predstavljen končni model, ki je bil kot najboljši izbran na podlagi rezultatov iz prvih dveh delov.

V četrtem delu (razdelek 4.1.4) so rezultati uvrščanja, ki so bili pridobljeni ob uporabi dveh metod za zmanjšanje vpliva neravnotežja, to sta bili metodi MDS in OAO.

V petem delu (razdelek 4.1.5) je predstavljena skladnost (konkordanca) mer za vrednotenje uvrščanja, ki smo jih uporabili za pridobitev rezultatov v prvih štirih delih.

#### 4.1.1 Predstavitev rezultatov modelov uvrščanja vrednotenih s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi

Prečno preverjanje z 10 pregibi smo izvedli tako, da smo množico podatkov razdelili na 10 delov. V vsakem od 10 korakov smo nato izvzeli eno od 10 množic in zgradili model uvrščanja na preostalih podatkih ter z njim napovedali pripadnost razredu za enote iz izvzete množice. Po 10 korakih smo za celotno množico podatkov dobili napovedano pripadnost razredu, ki smo jo uporabili za izračun mer za vrednotenje uvrščanja. Postopek smo nato 100 krat ponovili in iz pridobljenih vrednosti mer za vrednotenj uvrščanja izračunali povprečja.

Izmed obravnavanih metod uvrščanja so bili rezultati uvrščanja najslabši pri metodi CART (Tabela 4.1.1). Med vsemi modeli uvrščanja, ki so bili zgrajeni s CART, pa so bili rezultati najboljši ob izboru 50 spremenljivk z največjo varianco in uporabo 1. odvoda (Tabela 4.1.3).

##### Metoda

F-statistike je imela najvišje A-povprečje med obravnavanimi metodami najbližjega soseda. Če ob A-povprečju opazujemo še druge mere za vrednotenje napovedi uvrščanja (Tabela 4.1.6), vidimo, da pri kombinaciji metod 1-NN + 1. odvod + izbor 50 spremenljivk z največjo F-statistiko večina mer za vrednotenje uvrščanja doseže najvišjo vrednost. Vrednosti mer za vrednotenje uvrščanja pri obeh omenjenih kombinacijah so si zelo podobne. Na splošno je bila 1-NN uspešnejša od metod najbližjega soseda s 3 ali 5 sosedi (Tabeli B.1 in B.2). Z dodatnim povečanjem števila sosedov so se rezultati še poslabšali (rezultati niso prikazani).

Najboljše rezultate pri uvrščanju z metodo SVM (Tabeli 4.1.4 in 4.1.5) smo dobili v kombinaciji s predobdelavo SNV in zmanjšanjem dimenzije podatkov z metodo PCA, najslabše pa brez predobdelave in z izborom spremenljivk na podlagi največje variance. Pri uvrščanju z metodo SVM so bili rezultati modelov, kjer smo zmanjšali dimenzijo podatkov z metodo PCA na splošno boljši od tistih, kjer dimenzije nismo zmanjšali ali smo za zmanjšanje dimenzije uporabili izbor spremenljivk. Rezultati, ki smo jih dobili ob uporabi linearne jedrne funkcije in prilagojenega praga za uvrščanje glede na velikost razredov (Tabela 4.1.5), so veliko boljši kot rezultati, dobljeni z uporabo radialne jedrne funkcije brez prilagoditve praga za uvrščanje (Tabela 4.1.4). Glede na vrednosti mere G lahko trdimo, da se vpliv neravnotežja z uporabo linearne jedrne funkcije in prilagojene mere uvrščanja močno zmanjša.

Pri uvrščanju z metodo LDA (Tabela 4.1.2) se je za najmanj uspešno izkazala kombinacija: brez predobdelave + PCA; za najuspešnejšo pa: SNV predobdelava + brez izbora spremenljivk. Pri uvrščanju z metodo LDA so bili rezultati brez izbora spremenljivk na splošno boljši od tistih, kjer smo uporabili izbor spremenljivk ali zmanjšanje dimenzije podatkov z metodo PCA. Model s kombinacijo LDA + SNV predobdelava + brez izbora spremenljivk je bil tudi sicer izbran kot najboljši izmed vseh obravnavanih modelov.

Iz predstavljenih rezultatov smo opazili, da so vse obravnavane mere v večini primerov izbrale isti najboljši model. Rezultati o skladnosti mer na splošno, in ne le pri najboljšem modelu, so predstavljeni v razdelku 4.1.5. Največje neskladje med merami smo opazili

pri uvrščanju z metodo CART (Tabela 4.1.3), pri kateri so bili rezultati uvrščanja tudi najslabši.

Predobdelava	Zmanjšanje dim.	LDA	CART	R-SVM	L-SVM	1-NN	3-NN	5-NN
brez	brez	0,80	0,20	0,20	0,64	0,60	0,48	0,42
	varianca	0,63	0,12	0,05	0,36	0,39	0,28	0,27
	F-statistika	0,69	0,18	0,17	0,56	0,62	0,51	0,47
	PCA	0,38	0,24	0,35	0,61	0,54	0,44	0,40
kvant. norm.	brez	0,84	0,38	0,56	0,82	0,84	0,74	0,70
	varianca	0,79	0,36	0,44	0,84	0,81	0,70	0,65
	F-statistika	0,83	0,36	0,50	0,85	<b>0,87</b>	<b>0,78</b>	<b>0,74</b>
	PCA	0,76	0,31	0,63	0,85	0,79	0,70	0,66
1. odvod	brez	0,79	0,38	0,44	0,79	0,84	0,71	0,64
	varianca	0,70	<b>0,39</b>	0,33	0,74	0,82	0,69	0,61
	F-statistika	0,70	0,37	0,35	0,81	0,86	0,73	0,64
	PCA	0,65	0,33	0,56	0,77	0,83	0,70	0,64
SNV	brez	<b>0,88</b>	0,35	0,49	0,83	0,82	0,71	0,67
	varianca	0,84	0,34	0,39	0,81	0,78	0,69	0,63
	F-statistika	0,85	0,33	0,43	0,84	0,82	0,76	0,70
	PCA	0,79	0,31	<b>0,69</b>	<b>0,86</b>	0,80	0,71	0,66

Tabela 4.1.1: Povprečne vrednosti A-povprečja pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi. Najboljši rezultat za posamezno metodo uvrščanja (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,85	0,80	0,78	0,74	0,85	0,83	0,80	0,85	0,82	0,73	0,69	0,63	0,90	0,87	0,87	0,87
	varianca	0,69	0,63	0,20	0,50	0,68	0,66	0,61	0,68	0,66	0,54	0,48	0,29	0,81	0,75	0,75	0,75
	F-statistika	0,74	0,69	0,59	0,60	0,73	0,73	0,68	0,73	0,72	0,60	0,56	0,42	0,85	0,80	0,80	0,80
	PCA	0,49	0,38	0,00	0,07	0,47	0,37	0,35	0,48	0,48	0,30	0,26	-0,25	0,72	0,62	0,64	0,62
kvant. norm.	brez	0,90	0,84	0,68	0,76	0,89	0,87	0,85	0,89	0,86	0,80	0,76	0,72	0,94	0,92	0,92	0,92
	varianca	0,84	0,79	0,45	0,68	0,83	0,81	0,78	0,83	0,81	0,73	0,68	0,59	0,91	0,87	0,88	0,87
	F-statistika	0,86	0,83	0,80	0,76	0,86	0,86	0,83	0,86	0,85	0,77	0,74	0,68	0,92	0,89	0,90	0,89
	PCA	0,78	0,76	0,56	0,68	0,77	0,75	0,73	0,77	0,76	0,65	0,61	0,52	0,89	0,85	0,84	0,85
1. odvod	brez	0,85	0,79	0,77	0,73	0,84	0,82	0,78	0,84	0,81	0,72	0,67	0,61	0,90	0,87	0,87	0,87
	varianca	0,75	0,70	0,63	0,62	0,74	0,75	0,69	0,74	0,74	0,61	0,56	0,45	0,86	0,81	0,81	0,81
	F-statistika	0,76	0,70	0,57	0,60	0,75	0,77	0,70	0,75	0,75	0,62	0,59	0,47	0,86	0,80	0,81	0,80
	PCA	0,70	0,65	0,00	0,40	0,69	0,70	0,63	0,69	0,70	0,55	0,52	0,34	0,84	0,78	0,79	0,78
SNV	brez	<b>0,92</b>	<b>0,88</b>	<b>0,86</b>	<b>0,82</b>	<b>0,91</b>	<b>0,89</b>	<b>0,88</b>	<b>0,91</b>	<b>0,89</b>	<b>0,83</b>	<b>0,80</b>	<b>0,77</b>	<b>0,95</b>	<b>0,93</b>	<b>0,93</b>	<b>0,93</b>
	varianca	0,87	0,84	0,02	0,69	0,87	0,83	0,82	0,87	0,84	0,75	0,72	0,67	0,93	0,90	0,90	0,90
	F-statistika	0,88	0,85	0,83	0,78	0,88	0,86	0,84	0,88	0,86	0,78	0,74	0,71	0,93	0,91	0,91	0,91
	PCA	0,82	0,79	0,75	0,72	0,81	0,78	0,76	0,81	0,79	0,69	0,65	0,57	0,90	0,87	0,87	0,87

Tabela 4.1.2: **LDA (10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo LDA. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.



Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,36	0,20	0,00	0,02	0,33	0,18	0,17	0,33	0,37	0,15	0,12	-0,62	0,61	0,44	0,48	0,44
	varianca	0,21	0,12	0,00	0,01	0,16	0,08	0,08	0,17	0,39	0,07	0,05	-0,80	0,52	0,24	0,29	0,24
	F-statistika	0,34	0,18	0,00	0,02	0,30	0,17	0,15	0,30	0,39	0,13	0,10	-0,66	0,61	0,40	0,45	0,40
	PCA	0,43	0,24	0,00	0,02	0,40	0,20	0,21	0,40	0,41	0,18	0,15	-0,56	0,68	0,52	0,56	0,52
kvant. norm.	brez	<b>0,62</b>	0,38	0,00	0,05	<b>0,60</b>	0,34	0,35	<b>0,60</b>	<b>0,50</b>	<b>0,32</b>	<b>0,28</b>	-0,28	0,78	0,67	0,71	0,67
	varianca	0,61	0,36	0,00	0,04	0,59	0,33	0,33	0,59	<b>0,50</b>	0,30	0,27	-0,31	0,79	0,67	0,72	0,67
	F-statistika	0,60	0,36	0,00	0,04	0,58	0,32	0,32	0,58	0,48	0,28	0,25	-0,32	0,78	0,66	0,70	0,66
	PCA	0,52	0,31	0,00	0,04	0,50	0,27	0,28	0,50	0,44	0,24	0,20	-0,42	0,71	0,58	0,62	0,58
1. odvod	brez	0,59	0,38	0,00	<b>0,07</b>	0,57	<b>0,37</b>	<b>0,36</b>	0,58	0,48	<b>0,32</b>	<b>0,28</b>	<b>-0,25</b>	0,76	0,67	0,69	0,67
	varianca	0,61	<b>0,39</b>	0,00	0,05	0,59	0,36	<b>0,36</b>	0,59	<b>0,50</b>	0,31	<b>0,28</b>	-0,26	<b>0,80</b>	<b>0,70</b>	<b>0,73</b>	<b>0,70</b>
	F-statistika	0,60	0,37	0,00	0,06	0,58	0,36	0,35	0,58	0,47	<b>0,32</b>	0,27	-0,27	0,77	0,68	0,71	0,68
	PCA	0,55	0,33	0,00	0,05	0,53	0,32	0,30	0,53	0,44	0,27	0,22	-0,36	0,72	0,61	0,64	0,61
SNV	brez	0,58	0,35	0,00	0,05	0,56	0,33	0,33	0,56	0,47	0,30	0,25	-0,32	0,76	0,66	0,70	0,66
	varianca	0,56	0,34	0,00	0,04	0,54	0,31	0,31	0,54	0,47	0,28	0,23	-0,35	0,75	0,62	0,67	0,62
	F-statistika	0,56	0,33	0,00	0,04	0,53	0,29	0,30	0,54	0,45	0,26	0,22	-0,38	0,75	0,63	0,67	0,63
	PCA	0,52	0,31	0,00	0,04	0,50	0,28	0,28	0,50	0,44	0,24	0,20	-0,41	0,72	0,58	0,62	0,58

Tabela 4.1.3: **CART (10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo CART. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,41	0,20	0,00	0,01	0,37	0,20	0,17	0,38	0,46	0,14	0,12	-0,60	0,69	0,45	0,53	0,45
	varianca	0,13	0,05	0,00	0,00	0,06	0,02	0,03	0,06	0,40	0,02	0,02	-0,93	0,53	0,15	0,21	0,15
	F-statistika	0,36	0,17	0,00	0,01	0,32	0,18	0,14	0,33	0,48	0,11	0,10	-0,65	0,66	0,35	0,44	0,35
	PCA	0,55	0,35	0,00	0,05	0,53	0,38	0,34	0,54	0,53	0,29	0,25	-0,27	0,76	0,60	0,66	0,60
kvant. norm.	brez	0,72	0,56	0,00	0,13	0,71	0,61	0,55	0,71	0,70	0,49	0,47	0,16	0,87	0,74	0,81	0,74
	varianca	0,66	0,44	0,00	0,06	0,64	0,48	0,42	0,65	0,62	0,36	0,35	-0,08	0,84	0,69	0,76	0,69
	F-statistika	0,69	0,50	0,00	0,08	0,68	0,52	0,47	0,68	0,64	0,42	0,40	0,01	0,86	0,73	0,80	0,73
	PCA	0,79	0,62	0,00	0,23	0,78	0,68	0,62	0,78	0,73	0,56	0,54	0,30	0,89	0,81	0,85	0,81
1. odvod	brez	0,67	0,44	0,00	0,08	0,65	0,49	0,44	0,66	0,62	0,38	0,35	-0,06	0,82	0,68	0,74	0,68
	varianca	0,58	0,33	0,00	0,04	0,55	0,38	0,32	0,56	0,56	0,27	0,25	-0,29	0,79	0,58	0,67	0,58
	F-statistika	0,59	0,35	0,00	0,04	0,57	0,40	0,34	0,58	0,56	0,28	0,27	-0,26	0,80	0,62	0,70	0,62
	PCA	0,72	0,56	0,00	0,16	0,70	0,63	0,57	0,71	0,71	0,50	0,48	0,19	0,85	0,72	0,77	0,72
SNV	brez	0,70	0,49	0,00	0,08	0,68	0,50	0,47	0,68	0,64	0,42	0,40	-0,01	0,86	0,73	0,80	0,73
	varianca	0,62	0,39	0,00	0,04	0,60	0,38	0,36	0,60	0,57	0,31	0,29	-0,23	0,82	0,66	0,74	0,66
	F-statistika	0,66	0,43	0,00	0,05	0,65	0,43	0,40	0,65	0,60	0,35	0,33	-0,14	0,85	0,70	0,77	0,70
	PCA	<b>0,82</b>	<b>0,69</b>	0,00	<b>0,30</b>	<b>0,81</b>	<b>0,71</b>	<b>0,69</b>	<b>0,82</b>	<b>0,77</b>	<b>0,63</b>	<b>0,60</b>	<b>0,40</b>	<b>0,91</b>	<b>0,85</b>	<b>0,88</b>	<b>0,85</b>

Tabela 4.1.4: **R-SVM (10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo SVM z radialno jedrno funkcijo. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,61	0,64	0,16	0,51	0,60	0,62	0,58	0,60	0,64	0,49	0,45	0,26	0,78	0,72	0,71	0,72
	varianca	0,30	0,36	0,00	0,07	0,28	0,31	0,26	0,29	0,40	0,20	0,17	-0,33	0,63	0,51	0,52	0,51
	F-statistika	0,53	0,56	0,00	0,36	0,51	0,54	0,49	0,52	0,56	0,40	0,37	0,10	0,76	0,70	0,69	0,70
	PCA	0,58	0,61	0,08	0,45	0,57	0,56	0,54	0,57	0,60	0,46	0,41	0,16	0,77	0,73	0,71	0,73
kvant. norm.	brez	0,83	0,82	0,59	0,74	0,82	0,79	0,79	0,82	0,82	0,74	0,70	0,61	0,90	0,88	0,86	0,88
	varianca	0,82	0,84	0,70	0,77	0,81	0,79	0,80	0,82	0,82	0,73	0,69	0,63	0,90	0,87	0,86	0,87
	F-statistika	0,83	0,84	0,74	0,78	0,82	0,79	0,80	0,82	0,83	0,74	0,70	0,64	0,90	0,88	0,86	0,88
	PCA	0,83	<b>0,86</b>	0,80	0,79	0,83	0,80	0,81	0,83	<b>0,84</b>	0,75	<b>0,72</b>	0,66	0,90	0,88	<b>0,87</b>	0,88
1. odvod	brez	0,77	0,79	0,76	0,73	0,76	0,75	0,74	0,76	0,77	0,66	0,62	0,54	0,87	0,84	0,82	0,84
	varianca	0,73	0,74	0,71	0,68	0,72	0,71	0,69	0,72	0,73	0,60	0,56	0,46	0,85	0,82	0,80	0,82
	F-statistika	0,78	0,81	0,80	0,75	0,77	0,75	0,76	0,77	0,78	0,68	0,64	0,56	0,88	0,86	0,84	0,86
	PCA	0,74	0,77	0,70	0,70	0,73	0,71	0,71	0,74	0,75	0,63	0,58	0,48	0,85	0,82	0,80	0,82
SNV	brez	0,80	0,83	0,71	0,76	0,79	0,78	0,78	0,80	0,81	0,71	0,68	0,61	0,89	0,87	0,85	0,87
	varianca	0,77	0,81	0,69	0,73	0,76	0,74	0,74	0,76	0,77	0,66	0,62	0,54	0,87	0,85	0,83	0,85
	F-statistika	0,81	0,84	<b>0,82</b>	0,78	0,80	0,77	0,79	0,80	0,81	0,71	0,67	0,62	0,89	0,87	0,85	0,87
	PCA	<b>0,84</b>	<b>0,86</b>	0,76	<b>0,80</b>	<b>0,84</b>	<b>0,81</b>	<b>0,82</b>	<b>0,84</b>	<b>0,84</b>	<b>0,76</b>	<b>0,72</b>	<b>0,67</b>	<b>0,91</b>	<b>0,89</b>	<b>0,87</b>	<b>0,89</b>

Tabela 4.1.5: **L-SVM (10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo SVM, pri kateri je bila uporabljena linearna jedrna funkcija in je bil prag za uvrščanje prilagojen glede na velikost razredov. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,61	0,60	0,00	0,39	0,60	0,59	0,58	0,60	0,63	0,54	0,46	0,19	0,75	0,68	0,68	0,68
	varianca	0,41	0,39	0,00	0,14	0,38	0,40	0,38	0,38	0,47	0,35	0,28	-0,21	0,61	0,53	0,54	0,53
	F-statistika	0,68	0,62	0,00	0,40	0,66	0,62	0,61	0,66	0,66	0,56	0,48	0,24	0,79	0,72	0,72	0,72
	PCA	0,55	0,54	0,00	0,35	0,53	0,53	0,53	0,53	0,59	0,48	0,40	0,07	0,72	0,65	0,64	0,65
kvant. norm.	brez	0,87	0,84	0,80	0,75	0,86	0,84	0,83	0,86	0,85	0,80	0,75	0,68	0,92	0,89	0,89	0,89
	varianca	0,84	0,81	0,03	0,67	0,83	0,81	0,80	0,83	0,83	0,76	0,71	0,62	0,90	0,86	0,86	0,86
	F-statistika	<b>0,89</b>	<b>0,87</b>	<b>0,85</b>	<b>0,80</b>	0,88	<b>0,87</b>	<b>0,86</b>	0,88	<b>0,87</b>	<b>0,81</b>	<b>0,77</b>	<b>0,73</b>	0,92	0,90	0,90	0,90
	PCA	0,84	0,79	0,00	0,59	0,83	0,81	0,78	0,83	0,82	0,73	0,69	0,59	0,90	0,86	0,87	0,86
1. odvod	brez	0,87	0,84	0,82	0,77	0,87	0,84	0,83	0,87	0,85	0,78	0,74	0,68	0,92	0,89	0,89	0,89
	varianca	0,85	0,82	0,81	0,76	0,85	0,84	0,82	0,85	0,84	0,76	0,71	0,66	0,91	0,87	0,87	0,87
	F-statistika	<b>0,89</b>	0,86	<b>0,85</b>	<b>0,80</b>	<b>0,89</b>	0,86	0,85	<b>0,89</b>	<b>0,87</b>	0,80	<b>0,77</b>	<b>0,73</b>	<b>0,93</b>	<b>0,91</b>	<b>0,91</b>	<b>0,91</b>
	PCA	0,87	0,83	0,81	0,77	0,86	0,84	0,83	0,86	0,84	0,78	0,73	0,67	0,92	0,88	0,89	0,88
SNV	brez	0,86	0,82	0,01	0,66	0,85	0,83	0,81	0,85	0,84	0,76	0,72	0,65	0,91	0,87	0,88	0,87
	varianca	0,83	0,78	0,01	0,58	0,82	0,79	0,78	0,82	0,81	0,73	0,68	0,57	0,89	0,85	0,86	0,85
	F-statistika	0,86	0,82	0,78	0,73	0,85	0,82	0,81	0,85	0,83	0,76	0,71	0,64	0,91	0,88	0,88	0,88
	PCA	0,85	0,80	0,00	0,64	0,84	0,82	0,80	0,84	0,83	0,75	0,70	0,62	0,91	0,87	0,87	0,87

Tabela 4.1.6: **1-NN (10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo 1-NN. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

### 4.1.2 Predstavitev rezultatov modelov uvrščanja vrednotenih s 500 krat ponovljeno stratificirano razdelitvijo na učno in testno množico

Pri stratificirani razdelitvi na učno in testno množico smo podatke razdelili na 2 dela približno v razmerju 2 : 1, tako da so bili v obeh množicah predstavniki vseh razredov, pri čemer je bilo razmerje enot po razredih v obeh množicah približno enako. Na večji od obeh množic (učna množica) smo zgradili model uvrščanja, s katerim smo napovedali pripadnost razredu za enote v manjši množici (testna množica). Na podlagi teh napovedi smo model ovrednotili z merami za vrednotenje uvrščanja. Postopek smo nato 500 krat ponovili in iz pridobljenih vrednosti mer za vrednotenje uvrščanja izračunali povprečja.

Tako s 100 krat ponovljenim prečnim preverjanjem s pregibanjem kot s 500 krat ponovljeno razdelitvijo na učno in testno množico smo dobili podobne rezultate, le da so bile napovedne točnosti po razredih pri prečnem preverjanju nekoliko višje, kar je bilo pričakovano, saj smo pri prečnem preverjanju gradili modele uvrščanja na večjih učnih množicah.

Večjo razliko smo opazili pri uvrščanju z metodo R-SVM, kjer je bila pri prečnem preverjanju z 10 pregibi (Tabela 4.1.4) pri večini mer za vrednotenje uvrščanja kot najboljša izbrana kombinacija s predobdelavo SNV in uporabo PCA za zmanjšanje dimenzije podatkov. Pri 500 krat ponovljeni razdelitvi na učno in testno množico (Tabela 4.1.10) pa je bila pri večini mer za vrednotenje uvrščanja kot najboljša izbrana kombinacija s kvantilno normalizacijo kot predobdelavo in zmanjšanjem dimenzije podatkov z metodo PCA.

Predobdelava	Zmanjšanje dim.	LDA	CART	R-SVM	L-SVM	1-NN	3-NN	5-NN
brez	brez	0,78	0,20	0,17	0,51	0,55	0,41	0,35
	varianca	0,63	0,11	0,04	0,26	0,36	0,25	0,24
	F-statistika	0,68	0,17	0,14	0,45	0,58	0,45	0,40
	PCA	0,40	0,23	0,30	0,53	0,50	0,38	0,34
kvant. norm.	brez	0,81	<b>0,37</b>	0,45	0,73	0,81	0,71	0,61
	varianca	0,76	0,36	0,35	0,74	0,78	0,65	0,57
	F-statistika	0,81	0,35	0,42	0,76	<b>0,83</b>	<b>0,74</b>	<b>0,65</b>
	PCA	0,81	0,30	<b>0,66</b>	0,74	0,81	0,70	0,61
1. odvod	brez	0,78	0,36	0,37	0,70	0,80	0,65	0,56
	varianca	0,68	<b>0,37</b>	0,28	0,66	0,77	0,62	0,55
	F-statistika	0,68	0,36	0,32	0,69	0,82	0,66	0,57
	PCA	0,65	0,31	0,44	0,68	0,78	0,65	0,56
SNV	brez	<b>0,86</b>	0,34	0,40	0,72	0,79	0,67	0,59
	varianca	0,82	0,33	0,31	0,71	0,75	0,60	0,54
	F-statistika	0,84	0,32	0,38	0,73	0,80	0,72	0,64
	PCA	0,77	0,31	0,59	<b>0,78</b>	0,78	0,66	0,58

Tabela 4.1.7: Povprečne vrednosti A-povprečja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico. Najboljši rezultat za posamezno metodo uvrščanja (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,82	0,78	0,03	0,51	0,81	0,78	0,75	0,81	0,79	0,69	0,66	0,56	0,92	0,89	0,89	0,89
	varianca	0,68	0,63	0,00	0,31	0,67	0,66	0,60	0,67	0,67	0,53	0,50	0,28	0,86	0,80	0,82	0,80
	F-statistika	0,73	0,68	0,00	0,38	0,72	0,71	0,66	0,72	0,72	0,59	0,57	0,40	0,89	0,84	0,85	0,84
	PCA	0,49	0,40	0,00	0,08	0,47	0,39	0,37	0,48	0,51	0,32	0,29	-0,20	0,78	0,69	0,72	0,69
kvant. norm.	brez	0,86	0,81	0,03	0,54	0,86	0,82	0,80	0,86	0,83	0,74	0,72	0,63	0,94	0,91	0,92	0,91
	varianca	0,81	0,76	0,01	0,48	0,80	0,78	0,75	0,81	0,79	0,69	0,67	0,55	0,92	0,89	0,90	0,89
	F-statistika	0,84	0,80	0,06	0,55	0,84	0,83	0,79	0,84	0,83	0,73	0,72	0,63	0,94	0,90	0,91	0,90
	PCA	0,85	0,81	0,04	0,55	0,84	0,82	0,80	0,85	0,83	0,74	0,72	0,64	0,94	0,91	0,92	0,91
1. odvod	brez	0,81	0,78	0,01	0,50	0,81	0,77	0,75	0,81	0,79	0,68	0,66	0,55	0,92	0,89	0,89	0,89
	varianca	0,73	0,68	0,00	0,38	0,72	0,71	0,66	0,73	0,72	0,58	0,56	0,40	0,89	0,84	0,85	0,84
	F-statistika	0,74	0,68	0,00	0,36	0,73	0,72	0,66	0,73	0,73	0,60	0,58	0,40	0,89	0,84	0,85	0,84
	PCA	0,70	0,65	0,00	0,32	0,68	0,68	0,62	0,69	0,70	0,55	0,53	0,33	0,88	0,81	0,83	0,81
SNV	brez	<b>0,89</b>	<b>0,86</b>	0,10	<b>0,64</b>	<b>0,89</b>	<b>0,86</b>	<b>0,84</b>	<b>0,89</b>	<b>0,87</b>	<b>0,79</b>	<b>0,78</b>	<b>0,72</b>	<b>0,96</b>	<b>0,93</b>	<b>0,94</b>	<b>0,93</b>
	varianca	0,86	0,82	0,02	0,56	0,85	0,82	0,79	0,85	0,83	0,73	0,72	0,64	0,95	0,92	0,92	0,92
	F-statistika	0,87	0,84	<b>0,11</b>	0,62	0,87	0,85	0,82	0,87	0,85	0,77	0,75	0,69	0,95	0,92	0,93	0,92
	PCA	0,80	0,77	0,04	0,52	0,79	0,77	0,74	0,79	0,78	0,67	0,65	0,55	0,93	0,90	0,90	0,90

Tabela 4.1.8: **LDA (razdelitev)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo LDA. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,35	0,20	0,00	0,02	0,32	0,16	0,16	0,33	0,44	0,14	0,12	-0,63	0,71	0,49	0,57	0,49
	varianca	0,20	0,11	0,00	0,01	0,15	0,08	0,08	0,16	0,42	0,06	0,05	-0,82	0,61	0,31	0,40	0,31
	F-statistika	0,32	0,17	0,00	0,02	0,28	0,14	0,14	0,29	0,44	0,11	0,10	-0,69	0,69	0,44	0,54	0,44
	PCA	0,40	0,23	0,00	0,02	0,37	0,18	0,18	0,38	0,46	0,16	0,14	-0,59	0,74	0,54	0,62	0,54
kvant. norm.	brez	<b>0,58</b>	<b>0,37</b>	0,00	<b>0,05</b>	<b>0,56</b>	0,32	<b>0,33</b>	<b>0,57</b>	<b>0,54</b>	<b>0,29</b>	<b>0,27</b>	-0,31	0,83	0,69	0,76	0,69
	varianca	<b>0,58</b>	0,36	0,00	0,04	<b>0,56</b>	0,30	0,31	<b>0,57</b>	<b>0,54</b>	0,27	0,26	-0,34	<b>0,84</b>	0,68	0,76	0,68
	F-statistika	0,57	0,35	0,00	0,04	0,55	0,30	0,30	0,56	0,53	0,26	0,24	-0,35	0,83	0,68	0,76	0,68
	PCA	0,50	0,30	0,00	0,04	0,48	0,26	0,26	0,48	0,50	0,22	0,20	-0,44	0,79	0,62	0,70	0,62
1. odvod	brez	0,56	0,36	0,00	<b>0,05</b>	0,54	<b>0,34</b>	0,32	0,55	<b>0,54</b>	0,28	0,26	<b>-0,30</b>	0,83	0,68	0,75	0,68
	varianca	<b>0,58</b>	<b>0,37</b>	0,00	<b>0,05</b>	<b>0,56</b>	0,33	<b>0,33</b>	<b>0,57</b>	<b>0,54</b>	0,28	<b>0,27</b>	<b>-0,30</b>	<b>0,84</b>	<b>0,70</b>	<b>0,77</b>	<b>0,70</b>
	F-statistika	0,57	0,36	0,00	<b>0,05</b>	0,55	0,32	0,31	0,55	0,53	0,27	0,25	-0,32	<b>0,84</b>	<b>0,70</b>	0,76	<b>0,70</b>
	PCA	0,52	0,32	0,00	0,04	0,49	0,26	0,27	0,50	0,50	0,23	0,21	-0,42	0,80	0,64	0,72	0,64
SNV	brez	0,54	0,34	0,00	0,04	0,52	0,30	0,30	0,52	0,52	0,26	0,23	-0,36	0,81	0,66	0,73	0,66
	varianca	0,52	0,33	0,00	0,04	0,50	0,27	0,28	0,51	0,51	0,24	0,22	-0,40	0,81	0,65	0,72	0,65
	F-statistika	0,52	0,32	0,00	0,04	0,50	0,26	0,27	0,50	0,50	0,23	0,20	-0,42	0,81	0,65	0,72	0,65
	PCA	0,50	0,31	0,00	0,04	0,48	0,26	0,26	0,48	0,50	0,22	0,20	-0,43	0,79	0,63	0,70	0,63

Tabela 4.1.9: **CART (razdelitev)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo CART. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,36	0,17	0,00	0,01	0,32	0,14	0,14	0,33	0,45	0,11	0,10	-0,68	0,72	0,45	0,56	0,45
	varianca	0,12	0,04	0,00	0,00	0,04	0,02	0,02	0,05	0,44	0,02	0,01	-0,94	0,59	0,15	0,22	0,15
	F-statistika	0,32	0,14	0,00	0,01	0,27	0,12	0,11	0,29	0,46	0,08	0,08	-0,74	0,70	0,38	0,49	0,38
	PCA	0,49	0,30	0,00	0,04	0,46	0,31	0,27	0,47	0,52	0,23	0,21	-0,40	0,78	0,58	0,68	0,58
kvant. norm.	brez	0,65	0,45	0,00	0,08	0,63	0,45	0,42	0,64	0,61	0,36	0,35	-0,10	0,87	0,72	0,80	0,72
	varianca	0,58	0,35	0,00	0,04	0,55	0,32	0,31	0,56	0,56	0,27	0,26	-0,33	0,84	0,64	0,74	0,64
	F-statistika	0,62	0,42	0,00	0,06	0,60	0,39	0,38	0,61	0,59	0,33	0,32	-0,19	0,86	0,70	0,78	0,70
	PCA	<b>0,78</b>	<b>0,66</b>	0,00	<b>0,29</b>	<b>0,77</b>	<b>0,75</b>	<b>0,68</b>	<b>0,77</b>	<b>0,78</b>	<b>0,61</b>	<b>0,60</b>	<b>0,41</b>	0,90	0,80	0,85	0,80
1. odvod	brez	0,59	0,37	0,00	0,05	0,56	0,39	0,35	0,58	0,58	0,29	0,28	-0,24	0,83	0,64	0,73	0,64
	varianca	0,51	0,28	0,00	0,03	0,48	0,30	0,26	0,49	0,54	0,22	0,20	-0,42	0,79	0,56	0,67	0,56
	F-statistika	0,55	0,32	0,00	0,03	0,52	0,32	0,29	0,54	0,56	0,25	0,24	-0,36	0,82	0,61	0,72	0,61
	PCA	0,63	0,44	0,00	0,08	0,61	0,51	0,44	0,63	0,65	0,38	0,37	-0,05	0,83	0,64	0,74	0,64
SNV	brez	0,62	0,40	0,00	0,05	0,60	0,39	0,37	0,60	0,59	0,32	0,31	-0,21	0,86	0,69	0,78	0,69
	varianca	0,54	0,31	0,00	0,03	0,51	0,29	0,28	0,52	0,53	0,23	0,22	-0,40	0,82	0,62	0,72	0,62
	F-statistika	0,60	0,38	0,00	0,05	0,58	0,35	0,34	0,59	0,56	0,30	0,28	-0,27	0,85	0,69	0,78	0,69
	PCA	0,75	0,59	0,00	0,18	0,74	0,62	0,57	0,74	0,71	0,52	0,50	0,21	<b>0,91</b>	<b>0,81</b>	<b>0,86</b>	<b>0,81</b>

Tabela 4.1.10: **R-SVM (razdelitev)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo SVM z radialno jedrno funkcijo. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.



Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,48	0,51	0,00	0,17	0,47	0,50	0,44	0,47	0,54	0,36	0,34	0,01	0,79	0,73	0,73	0,73
	varianca	0,19	0,26	0,00	0,02	0,18	0,20	0,16	0,19	0,39	0,12	0,11	-0,54	0,67	0,49	0,55	0,49
	F-statistika	0,41	0,45	0,00	0,12	0,40	0,44	0,38	0,40	0,49	0,30	0,28	-0,11	0,77	0,70	0,72	0,70
	PCA	0,50	0,53	0,00	0,19	0,48	0,48	0,44	0,49	0,54	0,36	0,34	0,01	0,80	0,75	0,75	0,75
kvant. norm.	brez	0,71	0,73	0,02	0,44	0,70	0,72	0,68	0,71	0,74	0,61	0,59	0,46	<b>0,90</b>	0,86	0,86	0,86
	varianca	0,71	0,74	0,02	0,46	0,70	0,72	0,68	0,70	0,74	0,60	0,59	0,46	0,89	0,86	0,85	0,86
	F-statistika	0,74	0,76	0,01	0,48	0,73	0,73	0,71	0,73	0,75	0,63	0,61	0,49	<b>0,90</b>	0,87	0,86	0,87
	PCA	0,72	0,74	0,00	0,45	0,71	0,73	0,70	0,72	0,75	0,63	0,61	0,48	0,89	0,86	0,86	0,86
1. odvod	brez	0,67	0,70	0,00	0,40	0,66	0,68	0,63	0,66	0,69	0,55	0,53	0,37	0,88	0,84	0,83	0,84
	varianca	0,63	0,66	0,00	0,36	0,62	0,65	0,59	0,62	0,66	0,51	0,48	0,31	0,86	0,82	0,82	0,82
	F-statistika	0,66	0,69	0,00	0,39	0,65	0,67	0,63	0,66	0,69	0,54	0,52	0,36	0,88	0,84	0,84	0,84
	PCA	0,65	0,68	0,00	0,39	0,64	0,66	0,61	0,64	0,68	0,53	0,50	0,34	0,86	0,82	0,82	0,82
SNV	brez	0,70	0,72	0,00	0,42	0,69	0,70	0,67	0,69	0,72	0,59	0,57	0,43	0,89	0,85	0,85	0,85
	varianca	0,67	0,71	0,01	0,42	0,66	0,68	0,64	0,66	0,70	0,55	0,53	0,38	0,87	0,84	0,83	0,84
	F-statistika	0,69	0,73	0,00	0,44	0,68	0,70	0,66	0,69	0,72	0,58	0,56	0,43	0,88	0,85	0,84	0,85
	PCA	<b>0,76</b>	<b>0,78</b>	<b>0,06</b>	<b>0,52</b>	<b>0,75</b>	<b>0,75</b>	<b>0,72</b>	<b>0,75</b>	<b>0,77</b>	<b>0,64</b>	<b>0,63</b>	<b>0,53</b>	<b>0,90</b>	<b>0,88</b>	<b>0,87</b>	<b>0,88</b>

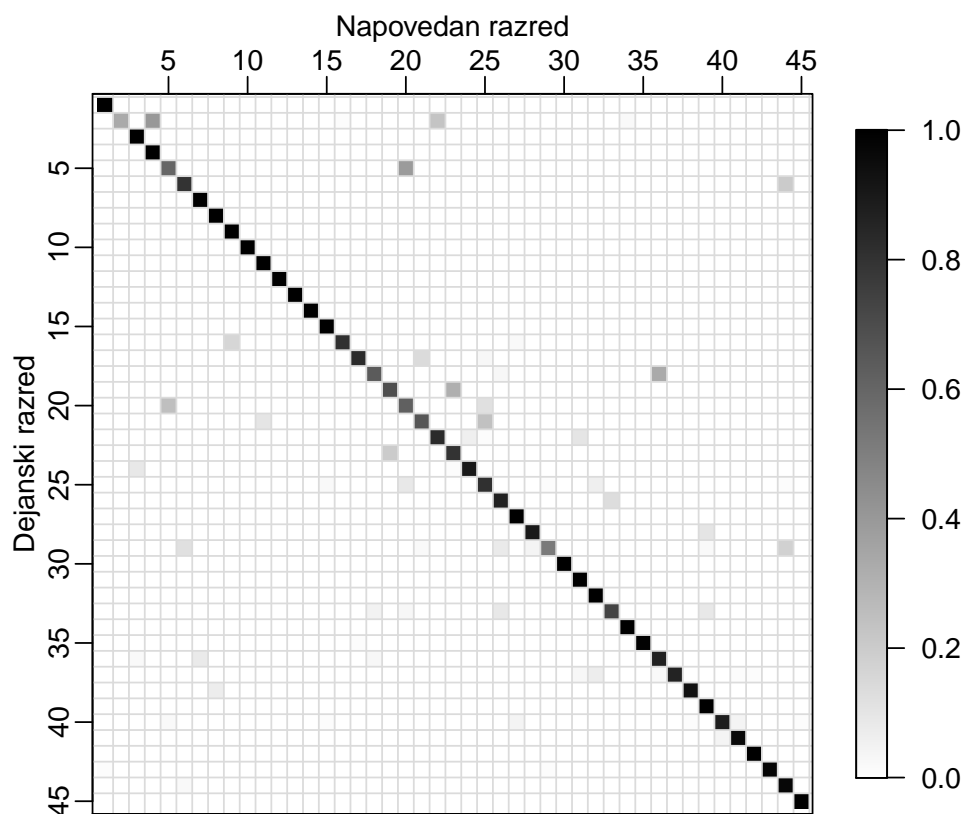
Tabela 4.1.11: **L-SVM (razdelitev)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo SVM, pri kateri je bila uporabljena linearna jedrna funkcija in je bil prag za uvrščanje prilagojen glede na velikost razredov. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,57	0,55	0,00	0,25	0,55	0,56	0,53	0,55	0,60	0,47	0,42	0,11	0,80	0,73	0,74	0,73
	varianca	0,37	0,36	0,00	0,09	0,34	0,37	0,34	0,35	0,45	0,30	0,26	-0,28	0,69	0,63	0,64	0,63
	F-statistika	0,62	0,58	0,00	0,27	0,61	0,59	0,56	0,61	0,62	0,50	0,45	0,16	0,82	0,76	0,77	0,76
	PCA	0,51	0,50	0,00	0,20	0,50	0,50	0,48	0,50	0,56	0,43	0,38	0,01	0,77	0,71	0,72	0,71
kvant. norm.	brez	0,84	0,81	0,02	0,55	0,83	0,83	0,80	0,83	0,83	0,74	0,72	0,64	0,93	0,90	0,90	0,90
	varianca	0,81	0,78	0,01	0,51	0,80	0,79	0,77	0,80	0,81	0,71	0,69	0,58	0,92	0,88	0,89	0,88
	F-statistika	<b>0,86</b>	<b>0,83</b>	0,09	<b>0,59</b>	<b>0,85</b>	<b>0,84</b>	<b>0,81</b>	<b>0,85</b>	<b>0,84</b>	<b>0,76</b>	<b>0,74</b>	<b>0,67</b>	<b>0,94</b>	0,90	0,91	0,90
	PCA	0,84	0,81	0,04	0,55	0,83	0,82	0,80	0,83	0,83	0,74	0,72	0,63	0,93	0,89	0,90	0,89
1. odvod	brez	0,83	0,80	0,07	0,55	0,83	0,81	0,78	0,83	0,82	0,72	0,70	0,61	0,93	0,90	0,91	0,90
	varianca	0,81	0,77	0,05	0,52	0,80	0,78	0,76	0,80	0,80	0,70	0,67	0,56	0,92	0,88	0,89	0,88
	F-statistika	0,85	0,82	<b>0,11</b>	0,58	<b>0,85</b>	0,83	0,80	<b>0,85</b>	0,83	0,75	0,73	0,65	<b>0,94</b>	<b>0,91</b>	<b>0,92</b>	<b>0,91</b>
	PCA	0,83	0,78	0,04	0,53	0,82	0,80	0,77	0,82	0,80	0,71	0,68	0,58	0,93	0,89	0,90	0,89
SNV	brez	0,82	0,79	0,00	0,51	0,82	0,80	0,77	0,82	0,81	0,72	0,69	0,59	0,92	0,88	0,89	0,88
	varianca	0,80	0,75	0,00	0,45	0,79	0,77	0,74	0,79	0,78	0,68	0,65	0,52	0,91	0,87	0,88	0,87
	F-statistika	0,83	0,80	0,05	0,54	0,83	0,81	0,78	0,83	0,82	0,72	0,70	0,60	0,93	0,89	0,90	0,89
	PCA	0,82	0,78	0,00	0,50	0,81	0,80	0,76	0,81	0,80	0,70	0,68	0,57	0,92	0,88	0,89	0,88

Tabela 4.1.12: **1-NN (razdelitev)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo 1-NN. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

### 4.1.3 Predstavitev končnega modela

Vrednosti mer za vrednotenje uvrščanja, pridobljene tako s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi kot 500 krat ponovljeno razdelitvijo na učno in testno množico, so bile najvišje za model uvrščanja, ki je bil zgrajen z metodo LDA na podatkih predobdelanih s SNV predobdelavo, pri katerih ni bila uporabljena nobena metoda izbora spremenljivk ali zmanjšanja dimenzije podatkov. Na Sliki 4.1.1 je grafično prikazana povprečna kontingenčna tabela izbranega modela, ki smo jo izračunali na podlagi napovedanih vrednosti, pridobljenih s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi. Prikazani so deleži, izračunani glede na dejansko število enot v razredih, povprečeni po 100 ponovitvah. Razredi so razporejeni od najmanjšega do največjega z enakimi oznakami kot v Tabeli 3.1.1. V Tabeli 4.1.13 pa so še bolj natančno predstavljene napačne napovedi izbranega modela uvrščanja.



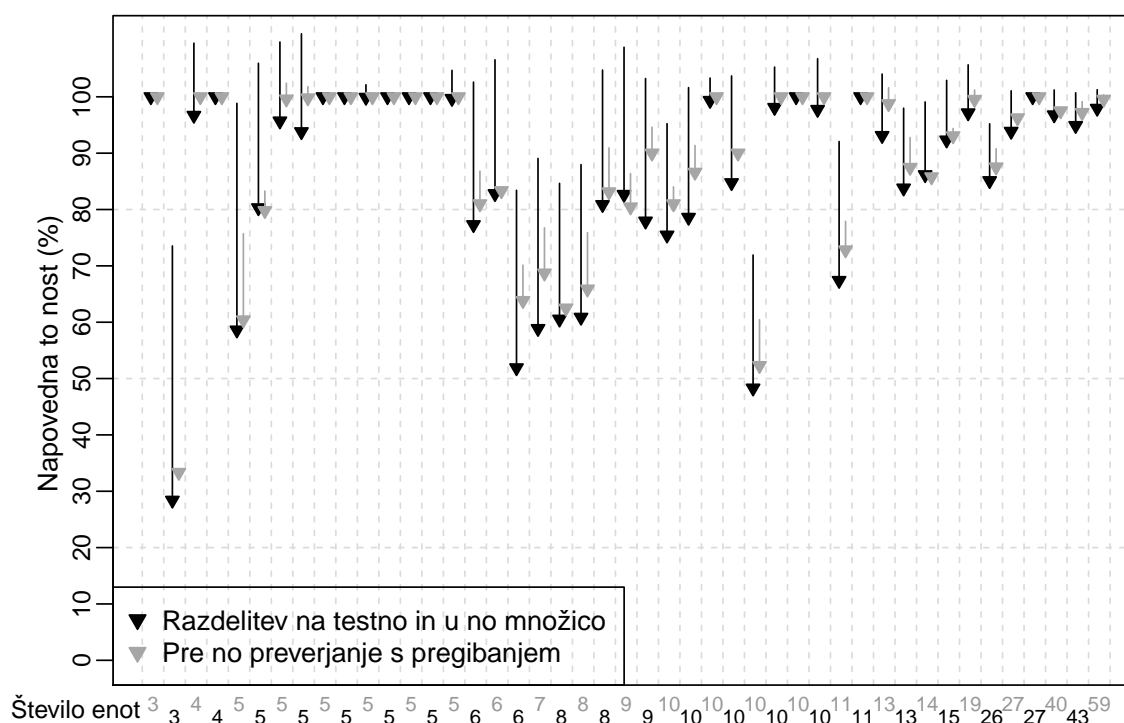
Slika 4.1.1: Grafično prikazana povprečna kontingenčna tabela, izračunana iz napovedi 100 krat ponovljenega prečnega preverjanja s pregibanjem za model uvrščanja, ki je bil zgrajen z metodo LDA na podatkih, predobdelanih s SNV, pri katerem ni bila uporabljena nobena metoda izbora spremenljivk ali zmanjšanja dimenzije podatkov. Prikazani so deleži, izračunani glede na dejansko število enot v razredih, povprečeni po 100 ponovitvah. Razredi so označeni od najmanjšega do največjega kot v Tabeli 3.1.1

Na Sliki 4.1.2 so prikazane povprečne napovedne točnosti po razredih s standardnimi odkloni za izbrani model, pridobljene tako s prečnim preverjanjem kot z razdelitvijo na učno in testno množico. Opazimo lahko, da so povprečne napovedne točnosti po razredih,

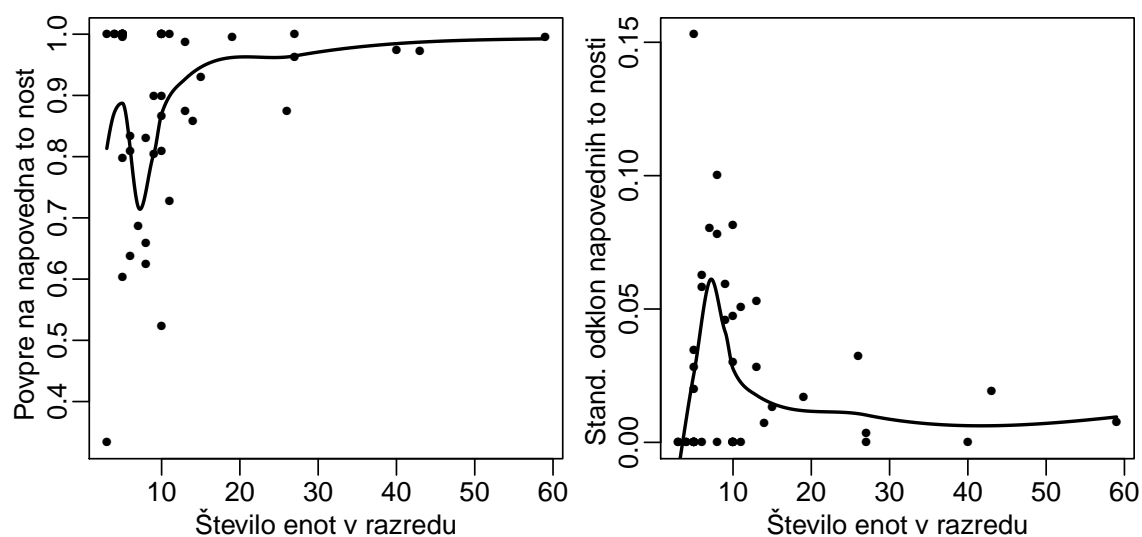
Dejanski razred		Napovedan razred		Delež
Oznaka Polimer		Oznaka Polimer		
2	Cellophane	4	Silk	0,40
5	Acrylonitrile-butadiene-styrene	20	Styrene-acrylonitrile	0,39
18	Leather	36	Bones, Teeth, Horns, Tortoiseshell, Ivory	0,33
19	Cellulose acetate propionate	23	Cellulose acetate butyrate	0,31
20	Styrene-acrylonitrile	5	Acrylonitrile-butadiene-styrene	0,25
21	Natural rubber	25	Styrene-butadiene	0,24
2	Cellophane	22	Poly(methyl methacrylate)	0,23
23	Cellulose acetate butyrate	19	Cellulose acetate propionate	0,20
6	Polyvinylchloride/polyvinylacetate	44	Poly(vinyl chloride)	0,20
29	Vinyl chloride-vinyl acetate-vinyl alcohol	44	Poly(vinyl chloride)	0,18
16	Polycarbonate	9	Polysulfone	0,16
17	Polytetrafluoroethylene	21	Natural rubber	0,14
26	Poly(propylene oxide)	33	Polyurethane (ether) (PUR)	0,13
29	Vinyl chloride-vinyl acetate-vinyl alcohol	6	Polyvinylchloride/polyvinylacetate	0,12
20	Styrene-acrylonitrile	25	Styrene-butadiene	0,12
21	Natural rubber	11	Polyisoprene	0,10
22	Poly(methyl methacrylate)	31	Urea-formaldehyde resin	0,10
28	Polyurethane (ester) (PUR)	39	Urethane elastomer thermoplastic	0,10
24	Cellulose acetate	3	Polyamide	0,09
25	Styrene-butadiene	20	Styrene-acrylonitrile	0,09
33	Polyurethane (ether) (PUR)	26	Poly(propylene oxide)	0,09
33	Polyurethane (ether) (PUR)	39	Urethane elastomer thermoplastic	0,09
36	Bones, Teeth, Horns, Tortoiseshell, Ivory	7	Casein formaldehyde	0,08
29	Vinyl chloride-vinyl acetate-vinyl alcohol	26	Poly(propylene oxide)	0,08
38	Polyamide	8	Melamine formaldehyde resin	0,07
37	Poly(ethylene terephthalate)	32	Epoxy resin	0,07
22	Poly(methyl methacrylate)	24	Cellulose acetate	0,06
25	Styrene-butadiene	32	Epoxy resin	0,06

Tabela 4.1.13: Napačno napovedani razredi, razporejeni po povprečnih deležih napačno napovedanih enot glede na velikost dejanskega razreda. Deleži so bili izračunani na podlagi rezultatov, pridobljenih s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi za izbrani model, ki je bil zgrajen z metodo uvrščanja LDA na podatkih predobdelanih s SNV predobdelavo, pri katerih ni bila uporabljena nobena metoda izbora spremenljivk ali zmanjšanja dimenzije podatkov. Prikazane so le napačne napovedi razredov, kjer je delež presegal 5 %. Oznake so enake kot v Tabeli 3.1.1.

pridobljene s prečnim preverjanjem na splošno višje standardni odkloni pa nižji od tistih, pridobljenih z razdelitvijo na učno in testno množico. Tako na Sliki 4.1.2 kot tudi na Sliki 4.1.3 lahko opazimo, da se povprečna napovedna točnost z velikostjo razreda zvišuje, standardni odklon pa znižuje, kar je tipična lastnost uvrščanja neuravnoteženih podatkov. Izjema so nekateri zelo majhni razredi, ki imajo 100 % napovedno točnost.



Slika 4.1.2: Povprečja in standardni odkloni napovednih točnosti po razredih, ki so bila izračunana iz 500 krat ponovljene razdelitve na učno in testno množico in 100 krat ponovljenega prečnega preverjanja z 10 pregibi za izbrani model, ki je bil zgrajen z metodo uvrščanja LDA na podatkih predobdelanih s SNV predobdelavo, pri katerih ni bila uporabljena nobena metoda izbora spremenljivk ali zmanjšanja dimenzije podatkov. Razredi so razporejeni od najmanjšega do največjega kot v Tabeli 3.1.1.



Slika 4.1.3: Povprečja in standardni odkloni napovednih točnosti po razredih v odvisnosti od števila enot v posameznem razredu. Povprečja in standardni odkloni so bili izračunani na podlagi rezultatov 100 krat ponovljenega prečnega preverjanja z 10 pregibi za izbrani model, ki je bil zgrajen z metodo uvrščanja LDA na podatkih predobdelanih s SNV predobdelavo, pri katerih ni bila uporabljena nobena metoda izbora spremenljivk ali zmanjšanja dimenzije podatkov. Glajena krivulja je bila izračunana s pomočjo funkcije *loess* iz programa **R** [140].

#### 4.1.4 Uvrščanje ob uporabi metod za zmanjšanje vpliva neravnotežja

##### 4.1.4.1 Uvrščanje ob uporabi večkratnega zmanjšanja večjih razredov (MDS)

Pri metodi večkratnega zmanjšanja večjega razreda smo iz učne množice z vzorčenjem brez ponavljanja iz vsakega razreda izbrali toliko enot, kot jih je v najmanjšem razredu, in tako tvorili novo učno množico, na kateri smo zgradili model uvrščanja. Postopek smo ponovili 100 krat. Enote iz testne množice smo napovedali z vsemi 100 modeli uvrščanja, končno pripadnost razredu smo določili z večinskim glasovanjem: enota je pripadala razredu, v katerega jo je uvrstilo največ modelov uvrščanja.

Če primerjamo rezultate uvrščanja z metodo LDA brez uporabe MDS (Tabela 4.1.2) z rezultati uvrščanja z metodo LDA ob uporabi MDS (Tabela 4.1.14) opazimo na splošno višje vrednosti mer za vrednotenje uvrščanja ob uporabi MDS. Tako kot brez uporabe metode MDS je bil tudi ob uporabi MDS kot najboljši izbran model uvrščanja na podatkih, predobdelanih s SNV in brez izbora spremenljivk. Prav pri tej kombinaciji metod (SNV + brez izbora spremenljivk + LDA) vrednosti mer za vrednotenje uvrščanja pri modelu z MDS niso bile višje od tistih pri modelu brez uporabe metod za zmanjšanje vpliva neravnotežja.

Pri uvrščanju z metodo CART opazimo, da so bili rezultati uvrščanja pri modelih brez zmanjšanja dimenzije podatkov ob uporabi metode MDS (Tabela 4.1.15) bistveno slabši od rezultatov uvrščanja brez uporabe MDS (Tabela 4.1.3). Na rezultate uvrščanja, kjer smo zmanjšali dimenzijo podatkov, uporaba metode MDS ni bistveno vplivala.

Pri uvrščanju z metodo R-SVM so bile pri modelu uvrščanja na podatkih, predobdelanih s kvantilno normalizacijo in z zmanjšano dimenzijo podatkov po metodi PCA, vrednosti mer za vrednotenje uvrščanja ob uporabi metode MDS (Tabela 4.1.16) višje kot brez uporabe metod za zmanjšanje vpliva neravnotežja (Tabela 4.1.4). Zato je bila ob uporabi MDS omenjena kombinacija (kvantilna normalizacija + PCA) med vsemi modeli uvrščanja z metodo R-SVM z večino mer za vrednotenje uvrščanja izbrana kot najboljša. Pri uvrščanju brez uporabe MDS je bila z večino mer za vrednotenje uvrščanja kot najboljša izbrana kombinacija: SNV predobdelava + PCA. Pri ostalih modelih uvrščanja z metodo R-SVM se rezultati uvrščanja ob uporabi MDS bistveno ne razlikujejo od rezultatov uvrščanja brez uporabe MDS.

Vrednosti mer za vrednotenje uvrščanja PA, K, MCC, CEN, RCI in NMI so bile pri uvrščanju z metodo L-SVM ob uporabi metode MDS višje kot brez uporabe MDS, ostale mere so bile višje le pri nekaterih kombinacijah metod. Najboljša izbrana kombinacija pri uvrščanju z metodo L-SVM in uporabo MDS je bila kvantilna normalizacija + brez izbora spremenljivk. Pri tej kombinaciji so bile vse mere za vrednotenje napovedi uvrščanja višje kot pri najboljši izbrani kombinaciji pri uvrščanju z metodo L-SVM brez uporabe MDS, ki je bila SNV predobdelava + zmanjšanje dimenzije podatkov z metodo PCA.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,82	0,84	0,83	0,78	0,82	0,79	0,80	0,82	0,82	0,72	0,68	0,64	0,90	0,88	0,87	0,88
	varianca	0,74	0,75	0,71	0,67	0,73	0,72	0,70	0,73	0,74	0,62	0,58	0,48	0,85	0,81	0,79	0,81
	F-statistika	0,80	0,80	0,74	0,72	0,79	0,77	0,76	0,80	0,79	0,68	0,64	0,57	0,89	0,86	0,85	0,86
	PCA	0,52	0,55	0,00	0,28	0,51	0,48	0,46	0,51	0,54	0,37	0,34	0,03	0,75	0,70	0,68	0,70
kvant. norm.	brez	0,85	0,85	0,84	0,79	0,84	0,84	0,83	0,85	0,85	0,77	0,74	0,69	0,92	0,89	0,89	0,89
	varianca	0,80	0,81	0,72	0,73	0,79	0,78	0,78	0,79	0,80	0,70	0,67	0,59	0,89	0,87	0,86	0,87
	F-statistika	0,85	0,87	0,85	0,80	0,84	0,84	0,83	0,85	0,85	0,76	0,74	0,71	0,92	0,90	0,89	0,90
	PCA	0,85	0,85	0,82	0,79	0,85	0,83	0,83	0,85	0,84	0,76	0,73	0,69	0,92	0,90	0,89	0,90
1. odvod	brez	0,82	0,84	0,81	0,77	0,81	0,78	0,79	0,81	0,81	0,71	0,68	0,62	0,90	0,88	0,86	0,88
	varianca	0,80	0,82	0,80	0,76	0,79	0,78	0,77	0,79	0,80	0,68	0,65	0,59	0,89	0,86	0,85	0,86
	F-statistika	0,82	0,82	0,80	0,75	0,81	0,79	0,78	0,81	0,81	0,70	0,67	0,60	0,90	0,87	0,86	0,87
	PCA	0,74	0,74	0,64	0,66	0,73	0,75	0,70	0,73	0,75	0,61	0,57	0,48	0,86	0,82	0,81	0,82
SNV	brez	<b>0,88</b>	<b>0,88</b>	<b>0,87</b>	<b>0,82</b>	<b>0,87</b>	<b>0,86</b>	<b>0,86</b>	<b>0,87</b>	<b>0,87</b>	<b>0,80</b>	<b>0,77</b>	<b>0,74</b>	<b>0,94</b>	<b>0,92</b>	<b>0,91</b>	<b>0,92</b>
	varianca	0,86	0,87	0,86	0,81	0,86	0,84	0,84	0,86	0,86	0,78	0,74	0,71	0,93	0,91	0,90	0,91
	F-statistika	0,87	<b>0,88</b>	0,86	0,81	<b>0,87</b>	0,85	0,85	<b>0,87</b>	0,86	0,78	0,76	0,73	0,93	<b>0,92</b>	<b>0,91</b>	<b>0,92</b>
	PCA	0,82	0,84	0,82	0,78	0,82	0,81	0,80	0,82	0,82	0,72	0,69	0,65	0,92	0,90	0,88	0,90

Tabela 4.1.14: **LDA (MDS, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo LDA ob uporabi metode večkratnega zmanjšanja večjih razredov. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.



Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,29	0,13	0,00	0,01	0,25	0,10	0,10	0,25	0,41	0,08	0,07	-0,78	0,58	0,30	0,36	0,30
	varianca	0,21	0,12	0,00	0,01	0,16	0,08	0,08	0,17	0,39	0,07	0,05	-0,80	0,52	0,24	0,29	0,24
	F-statistika	0,33	0,18	0,00	0,01	0,29	0,16	0,15	0,30	0,39	0,12	0,10	-0,66	0,60	0,39	0,45	0,39
	PCA	0,43	0,24	0,00	0,02	0,40	0,20	0,20	0,40	0,41	0,18	0,14	-0,56	0,68	0,52	0,56	0,52
kvant. norm.	brez	0,34	0,10	0,00	0,00	0,29	0,07	0,08	0,34	0,49	0,07	0,07	-0,83	0,68	0,28	0,40	0,28
	varianca	<b>0,61</b>	0,36	0,00	0,04	<b>0,59</b>	0,33	0,33	0,59	0,49	0,30	0,27	-0,31	0,79	0,68	0,72	0,68
	F-statistika	0,60	0,36	0,00	0,04	0,57	0,33	0,32	0,58	0,48	0,29	0,25	-0,31	0,77	0,64	0,69	0,64
	PCA	0,55	0,32	0,00	0,04	0,52	0,31	0,30	0,53	0,45	0,27	0,23	-0,37	0,73	0,61	0,64	0,61
1. odvod	brez	0,33	0,11	0,00	0,00	0,28	0,09	0,09	0,33	0,48	0,08	0,07	-0,80	0,66	0,26	0,37	0,26
	varianca	<b>0,61</b>	<b>0,39</b>	0,00	0,05	<b>0,59</b>	<b>0,36</b>	<b>0,36</b>	<b>0,60</b>	<b>0,50</b>	<b>0,32</b>	<b>0,28</b>	<b>-0,26</b>	<b>0,80</b>	<b>0,70</b>	<b>0,73</b>	<b>0,70</b>
	F-statistika	0,60	0,37	0,00	<b>0,06</b>	0,58	<b>0,36</b>	0,35	0,58	0,48	<b>0,32</b>	0,27	<b>-0,26</b>	0,77	0,68	0,71	0,68
	PCA	0,55	0,33	0,00	0,05	0,53	0,31	0,30	0,53	0,44	0,26	0,22	-0,36	0,73	0,61	0,64	0,61
SNV	brez	0,29	0,09	0,00	0,00	0,25	0,07	0,07	0,27	0,46	0,07	0,06	-0,84	0,63	0,26	0,35	0,26
	varianca	0,56	0,34	0,00	0,04	0,54	0,31	0,31	0,54	0,47	0,27	0,23	-0,35	0,75	0,62	0,67	0,62
	F-statistika	0,56	0,33	0,00	0,04	0,54	0,29	0,30	0,54	0,45	0,26	0,22	-0,38	0,76	0,63	0,67	0,63
	PCA	0,52	0,31	0,00	0,04	0,50	0,28	0,28	0,50	0,44	0,25	0,20	-0,41	0,72	0,59	0,62	0,59

Tabela 4.1.15: **CART (MDS, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo CART ob uporabi metode večkratnega zmanjšanja večjih razredov. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,41	0,20	0,00	0,01	0,37	0,20	0,17	0,38	0,46	0,14	0,12	-0,60	0,70	0,45	0,53	0,45
	varianca	0,13	0,05	0,00	0,00	0,06	0,02	0,03	0,06	0,40	0,02	0,02	-0,93	0,53	0,16	0,21	0,16
	F-statistika	0,36	0,17	0,00	0,01	0,32	0,18	0,14	0,33	0,48	0,11	0,10	-0,65	0,66	0,35	0,44	0,35
	PCA	0,56	0,35	0,00	0,05	0,53	0,38	0,34	0,54	0,54	0,29	0,26	-0,27	0,76	0,60	0,66	0,60
kvant. norm.	brez	0,72	0,56	0,00	0,13	0,71	0,61	0,55	0,71	0,70	0,48	0,47	0,16	0,87	0,74	0,80	0,74
	varianca	0,66	0,44	0,00	0,06	0,64	0,48	0,43	0,65	0,62	0,36	0,35	-0,08	0,84	0,69	0,76	0,69
	F-statistika	0,69	0,48	0,00	0,08	0,67	0,51	0,46	0,67	0,63	0,40	0,38	-0,01	0,86	0,72	0,79	0,72
	PCA	<b>0,84</b>	<b>0,74</b>	0,00	<b>0,40</b>	<b>0,83</b>	<b>0,80</b>	<b>0,76</b>	<b>0,83</b>	<b>0,82</b>	<b>0,70</b>	<b>0,68</b>	<b>0,54</b>	0,90	0,83	0,86	0,83
1. odvod	brez	0,67	0,44	0,00	0,08	0,65	0,50	0,44	0,66	0,62	0,38	0,35	-0,06	0,82	0,68	0,74	0,68
	varianca	0,58	0,33	0,00	0,04	0,55	0,38	0,32	0,56	0,56	0,27	0,25	-0,29	0,78	0,58	0,67	0,58
	F-statistika	0,59	0,35	0,00	0,04	0,57	0,40	0,34	0,58	0,56	0,28	0,27	-0,26	0,81	0,62	0,70	0,62
	PCA	0,72	0,56	0,00	0,16	0,70	0,63	0,56	0,71	0,70	0,50	0,48	0,19	0,85	0,72	0,77	0,72
SNV	brez	0,70	0,49	0,00	0,08	0,68	0,50	0,47	0,69	0,64	0,42	0,40	-0,01	0,86	0,74	0,80	0,74
	varianca	0,62	0,39	0,00	0,04	0,60	0,38	0,36	0,60	0,57	0,31	0,29	-0,23	0,82	0,66	0,74	0,66
	F-statistika	0,66	0,43	0,00	0,05	0,65	0,43	0,40	0,65	0,60	0,35	0,33	-0,14	0,85	0,70	0,77	0,70
	PCA	0,82	0,69	0,00	0,30	0,81	0,71	0,69	0,82	0,77	0,63	0,60	0,40	<b>0,91</b>	<b>0,85</b>	<b>0,88</b>	<b>0,85</b>

Tabela 4.1.16: **R-SVM (MDS, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo SVM, pri kateri je bila uporabljena radialna jedrna funkcija, ob uporabi metode večkratnega zmanjšanja večjih razredov. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,79	0,71	0,00	0,52	0,78	0,74	0,70	0,78	0,75	0,65	0,59	0,45	0,87	0,82	0,83	0,82
	varianca	0,45	0,25	0,00	0,02	0,41	0,23	0,22	0,42	0,47	0,18	0,16	-0,52	0,70	0,48	0,55	0,48
	F-statistika	0,64	0,45	0,00	0,10	0,63	0,47	0,44	0,63	0,58	0,39	0,35	-0,08	0,80	0,69	0,73	0,69
	PCA	0,63	0,46	0,00	0,11	0,61	0,47	0,44	0,62	0,57	0,39	0,35	-0,07	0,80	0,69	0,73	0,69
kvant. norm.	brez	<b>0,90</b>	0,85	<b>0,82</b>	0,77	<b>0,89</b>	0,86	<b>0,85</b>	<b>0,89</b>	0,86	<b>0,81</b>	<b>0,77</b>	0,71	<b>0,94</b>	<b>0,91</b>	<b>0,91</b>	<b>0,91</b>
	varianca	0,87	0,83	0,70	0,74	0,87	0,85	0,82	0,87	0,85	0,78	0,74	0,67	0,92	0,89	0,89	0,89
	F-statistika	0,87	0,82	0,63	0,73	0,87	0,83	0,82	0,87	0,84	0,77	0,72	0,66	0,92	0,89	0,90	0,89
	PCA	0,89	0,85	<b>0,82</b>	<b>0,78</b>	0,88	0,86	0,84	0,88	0,86	0,80	0,76	0,71	0,93	0,90	0,90	0,90
1. odvod	brez	0,87	0,81	0,78	0,74	0,86	0,82	0,80	0,86	0,83	0,74	0,70	0,63	0,92	0,88	0,88	0,88
	varianca	0,80	0,69	0,00	0,41	0,79	0,75	0,70	0,79	0,75	0,65	0,61	0,44	0,89	0,84	0,85	0,84
	F-statistika	0,84	0,74	0,00	0,45	0,83	0,76	0,73	0,83	0,78	0,68	0,64	0,49	0,91	0,87	0,88	0,87
	PCA	0,84	0,78	0,12	0,62	0,83	0,79	0,77	0,84	0,80	0,71	0,66	0,56	0,90	0,87	0,87	0,87
SNV	brez	0,89	<b>0,86</b>	0,80	<b>0,78</b>	<b>0,89</b>	<b>0,87</b>	<b>0,85</b>	<b>0,89</b>	<b>0,87</b>	0,80	<b>0,77</b>	<b>0,73</b>	0,93	0,90	<b>0,91</b>	0,90
	varianca	0,85	0,79	0,74	0,71	0,85	0,81	0,79	0,85	0,81	0,73	0,68	0,60	0,91	0,87	0,88	0,87
	F-statistika	<b>0,90</b>	0,85	0,00	0,70	<b>0,89</b>	0,86	0,84	<b>0,89</b>	<b>0,87</b>	0,79	0,76	0,71	<b>0,94</b>	<b>0,91</b>	<b>0,91</b>	<b>0,91</b>
	PCA	0,89	0,85	0,78	0,77	0,88	0,85	0,84	0,88	0,86	0,79	0,75	0,70	<b>0,94</b>	<b>0,91</b>	<b>0,91</b>	<b>0,91</b>

Tabela 4.1.17: **L-SVM (MDS, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo SVM, pri kateri je bila uporabljena linearna jedrna funkcija, ob uporabi metode večkratnega zmanjšanja večjih razredov. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

#### 4.1.4.2 Uvrščanje ob uporabi metode OAO

Pri metodi OAO zgradimo model uvrščanja za vsak par razredov. Novo enoto nato napovemo z vsemi zgrajenimi modeli. Končno pripadnost razredu nato določimo z večinskim glasovanjem. Metodo OAO smo uporabili le v kombinaciji s predobdelavo SNV in metodami za uvrščanje LDA, CART in R-SVM.

Pri rezultatih uvrščanja z metodo LDA ob uporabi metode OAO (Tabela 4.1.18) in brez uporabe OAO (Tabela 4.1.2) nismo opazili večjih razlik, edino rezultati uvrščanja z uporabo metode PCA za zmanjšanje dimenzije podatkov so bili boljši, ko je bila uporabljena OAO.

Rezultati uvrščanja z metodo CART pri uporabi metode OAO (Tabela 4.1.19) pa so veliko boljši kot brez uporabe metode OAO (Tabela 4.1.3) ali z uporabo metode MDS za zmanjšanje vpliva neravnotežja (Tabela 4.1.15). Podobno opazimo pri uvrščanju z metodo R-SVM, če primerjamo rezultate prikazane v Tabelah 4.1.20, 4.1.16 in 4.1.4, kar je nekoliko nenavadno, saj je metoda OAO že vgrajena v funkcijo *svm*, ki je bila uporabljena za gradnjo modelov uvrščanja z metodo R-SVM.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
SNV	brez	<b>0,89</b>	0,84	0,62	0,74	<b>0,88</b>	<b>0,88</b>	<b>0,84</b>	<b>0,88</b>	<b>0,86</b>	<b>0,78</b>	<b>0,75</b>	<b>0,71</b>	<b>0,93</b>	<b>0,90</b>	<b>0,91</b>	<b>0,90</b>
	varianca	0,86	0,83	0,59	0,72	0,86	0,84	0,81	0,86	0,84	0,74	0,72	0,67	0,92	0,89	0,89	0,89
	F-statistika	0,88	<b>0,85</b>	<b>0,77</b>	<b>0,76</b>	0,87	0,86	<b>0,84</b>	0,87	<b>0,86</b>	<b>0,78</b>	<b>0,75</b>	<b>0,71</b>	<b>0,93</b>	<b>0,90</b>	0,90	<b>0,90</b>
	PCA	<b>0,89</b>	0,83	0,57	0,72	<b>0,88</b>	<b>0,88</b>	<b>0,84</b>	<b>0,88</b>	<b>0,86</b>	<b>0,78</b>	<b>0,75</b>	<b>0,71</b>	<b>0,93</b>	0,89	0,90	0,89

Tabela 4.1.18: **LDA (OAO, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo LDA, ob uporabi metode vsak proti vsakem. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
SNV	brez	0,74	0,64	<b>0,60</b>	0,57	0,73	0,74	0,67	0,73	0,72	0,59	0,53	0,39	0,84	0,77	0,78	0,77
	varianca	0,76	0,67	0,08	0,55	0,74	0,75	0,69	0,74	0,73	0,63	0,56	0,42	0,84	0,78	0,79	0,78
	F-statistika	0,78	0,71	0,32	0,61	0,78	<b>0,78</b>	0,72	0,78	0,76	0,67	0,60	0,48	0,86	0,80	0,81	0,80
	PCA	<b>0,80</b>	<b>0,73</b>	0,29	<b>0,62</b>	<b>0,80</b>	<b>0,78</b>	<b>0,74</b>	<b>0,80</b>	<b>0,78</b>	<b>0,69</b>	<b>0,63</b>	<b>0,52</b>	<b>0,88</b>	<b>0,82</b>	<b>0,83</b>	<b>0,82</b>

Tabela 4.1.19: **CART (OAO, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo CART, ob uporabi metode vsak proti vsakem. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
SNV	brez	0,83	0,73	0,00	0,44	0,82	0,80	0,75	0,82	0,82	0,69	0,66	0,53	0,90	0,83	0,86	0,83
	varianca	0,78	0,66	0,00	0,35	0,77	0,72	0,67	0,77	0,75	0,61	0,57	0,39	0,87	0,79	0,82	0,79
	F-statistika	0,82	0,73	0,00	0,45	0,81	0,79	0,74	0,81	0,80	0,69	0,65	0,52	0,89	0,83	0,85	0,83
	PCA	<b>0,86</b>	<b>0,77</b>	0,00	<b>0,51</b>	<b>0,86</b>	<b>0,86</b>	<b>0,79</b>	<b>0,86</b>	<b>0,85</b>	<b>0,74</b>	<b>0,72</b>	<b>0,63</b>	<b>0,92</b>	<b>0,86</b>	<b>0,88</b>	<b>0,86</b>

Tabela 4.1.20: **R-SVM (OAO, 10-CV)**. Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo SVM, pri kateri je bila uporabljena radialna jedrna funkcija, ob uporabi metode vsak proti vsakem. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

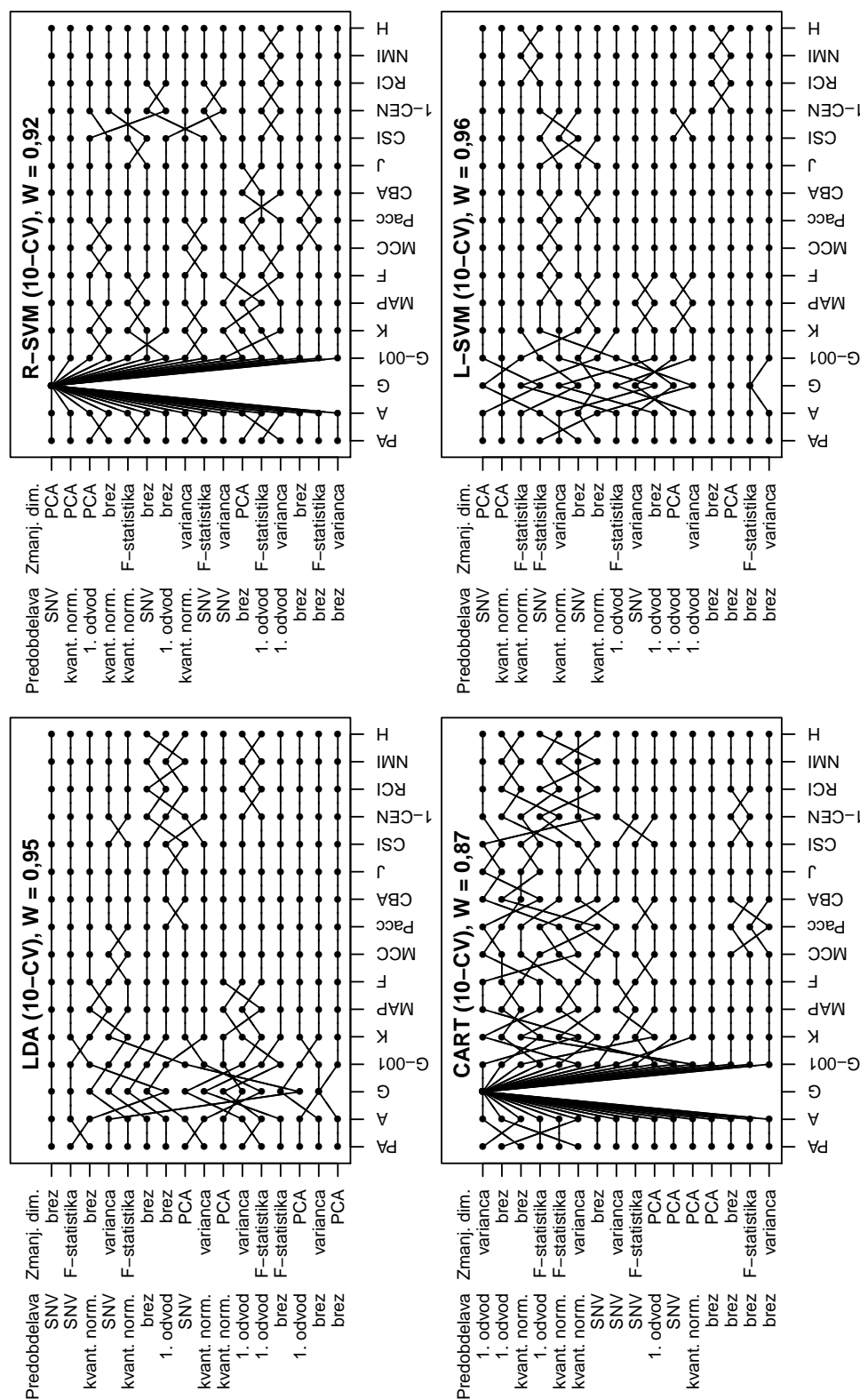
#### 4.1.5 Skladnost mer za vrednotenje uvrščanja

V razdelkih 4.1.1 in 4.1.2 smo predstavili rezultate 300 modelov uvrščanja. Vse modele smo ovrednotili s 16 merami za vrednotenje uvrščanja, pri čemer smo pri 188 uporabili 100 krat ponovljeno prečno preverjanje z 10 pregibi, pri 122 pa 500 krat ponovljeno razdelitev na testno in učno množico. V tabelah, kjer so bile prikazane mere za vrednotenje uvrščanja (kot npr. Tabele 4.1.2, 4.1.3, 4.1.4, itd.) smo lahko opazili, da so mere večinoma enotno izbrale najboljši model. V tem razdelku bomo predstavili, kako skladne so obravnavane mere na splošno.

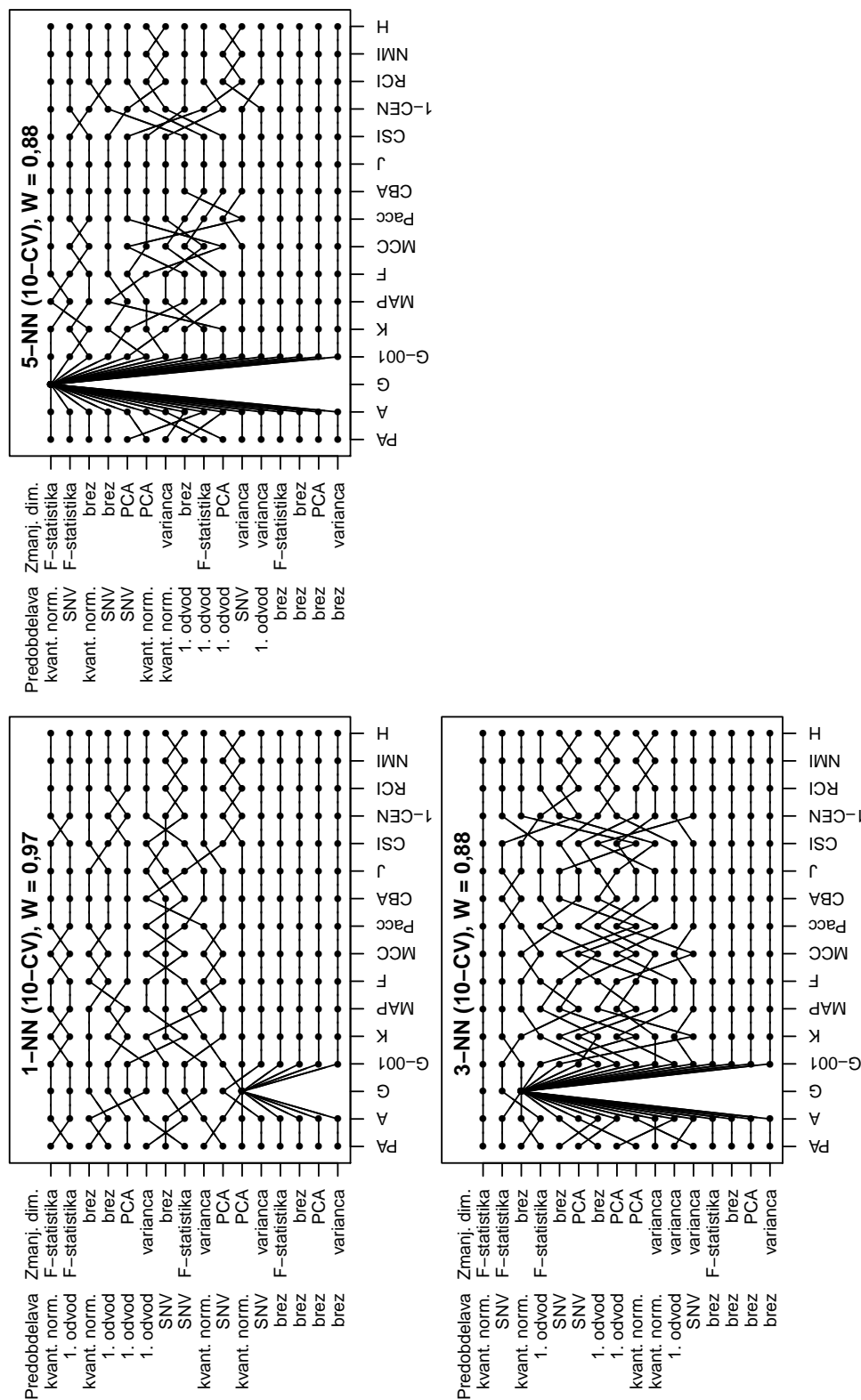
Skladnost mer smo najprej prikazali po skupinah s 16 modeli, kot so bile prikazane tudi v tabelah v razdelkih 4.1.1 in 4.1.2 (Tabele 4.1.2, 4.1.3, 4.1.4, itd.). Na Slikah 4.1.4 in 4.1.5 je prikaz konkordance z rangi na vzporednih oseh (opis metode je v razdelku 3.3.5) za modele uvrščanja, pri katerih ni bila uporabljena nobena metoda za zmanjšanje vpliva neravnotežja in so bile mere uvrščanja izračunane s pomočjo 100 krat ponovljenega prečnega preverjanja z 10 pregibi. Konkordanca z rangi na vzporednih oseh za modele uvrščanja, ki so bili ovrednoteni s pomočjo 500 krat ponovljene razdelitve na testno in učno množico, in za modele, kjer je bila za zmanjšanje vpliva neravnotežja uporabljena metoda MDS pa je prikazana v dodatku C. Na slikah prikaza konkordance z rangi na vzporednih oseh več križanj med črtami predstavlja manjšo konkordanco in obratno, manj križanj večjo konkordanco. Opazimo lahko, da se je mera G najslabše skladala z ostalimi, kar je bilo še toliko bolj izrazito pri skupini modelov, ki so na splošno imeli nižje vrednosti mer za vrednotenje uvrščanja. Npr. na Sliki 4.1.4 pri modelih na grafu LDA (10-CV) opazimo večjo skladnost mere G z ostalimi merami kot na grafu CART (10-CV). V Tabelah 4.1.2 in 4.1.3 pa smo videli, da so modeli zgrajeni z metodo CART na splošno slabši, kot modeli zgrajeni z metodo LDA. Na Slikah 4.1.4 in 4.1.5 pri grafu CART (10-CV) opazimo, da je skladnost tudi med drugimi merami slabša kot pri drugih metodah uvrščanja.

Na Sliki 4.1.6 so grafično prikazani Kendalovi koeficienti korelacije rangov  $\tau$  za vse pare obravnavanih mer, ki so bili izračunani na vrednostih mer pri vseh 300 obravnavanih modelih. Vse vrednosti  $\tau$  so bile pozitivne, kar kaže na pozitivno usklajenost vseh parov mer za vrednotenje uvrščanja. To je bilo pričakovano, saj so pri vseh obravnavanih merah nižje vrednosti predstavljale slabši model, višje vrednosti pa boljši model. Izjema je bila le mera CEN, pri kateri pa smo uporabili vrednosti 1-CEN. Na sliki izstopajo vrednosti korelacijskih koeficientov pri meri G, ki so zelo nizke v primerjavi z ostalimi, kar kaže na slabo usklajenost mere G z ostalimi merami. Mnogo bolje od mere G je z ostalimi merami usklajena mera G-001, čeprav lahko pri njej še vedno opazimo nekaj nizkih vrednosti koeficienta  $\tau$  (npr. pri 1-CEN in NMI). Ostale mere so med seboj sorazmerno dobro usklajene z vrednostmi koeficienta  $\tau > 0,7$ . Najboljšo usklajenost opazimo med merami: PA, K in MCC ter med merama H in RCI ( $\tau > 0,99$ ).

Konkordanca obravnavanih mer pri vseh 300 modelih je prikazana tudi s konkordančnim mehurčnim diagramom (metoda je opisana v razdelku 3.3.5), ki je prikazan na Sliki 4.1.7. Pri konkordančnem mehurčnem diagramu so mere med seboj toliko bolj usklajene, kolikor bolj se krogi prilegajo glavni diagonalni. Izstopajoči krogi v navpični smeri, ki jih na sliki opazimo približno pri rangju 190, so posledica mere G, za katero smo že na prejšnjih slikah ugotovili, da izstopa od vseh ostalih. Za ostale kroge v konkordančnem mehurčnem diagramu lahko rečemo, da se z nekaj izjemami sorazmerno dobro prilegajo glavni diagonalni, kar se sklada z visoko vrednostjo skupnega koeficienta konkordance  $W = 0,95$ .

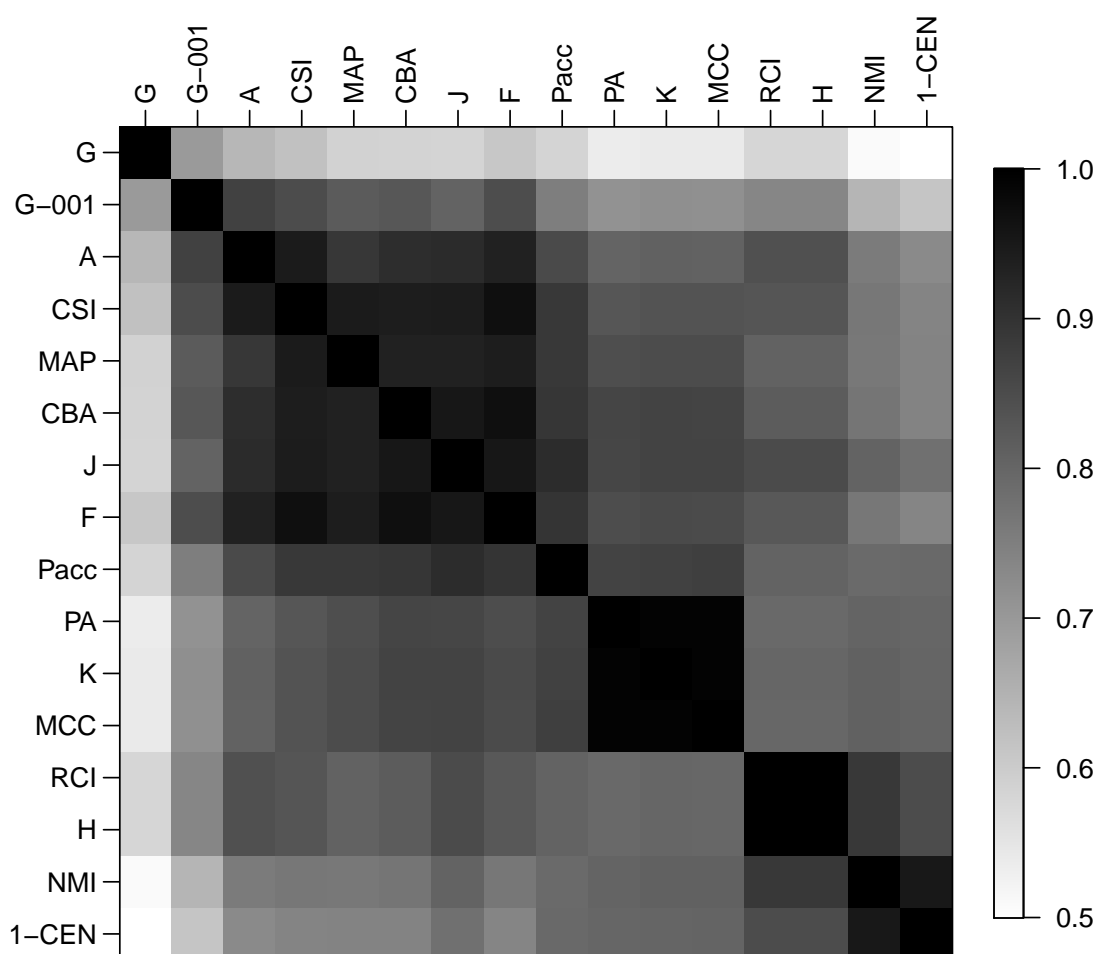


Slika 4.1.4: Prikaz konkordance 16 mer za vrednotenje uvrščanja s Kendallovim koeficientom konkordance ( $W$ ) in prikazom rangov na vzporednih oseh za modele uvrščanja zgrajene z metodami LDA, CART, R-SVM in L-SVM ter ovrednotene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi. Modeli so za vsako od metod uvrščanja razporejeni od najboljšega (zgoraj) do najslabšega (spodaj) glede na mero A-povprečje.

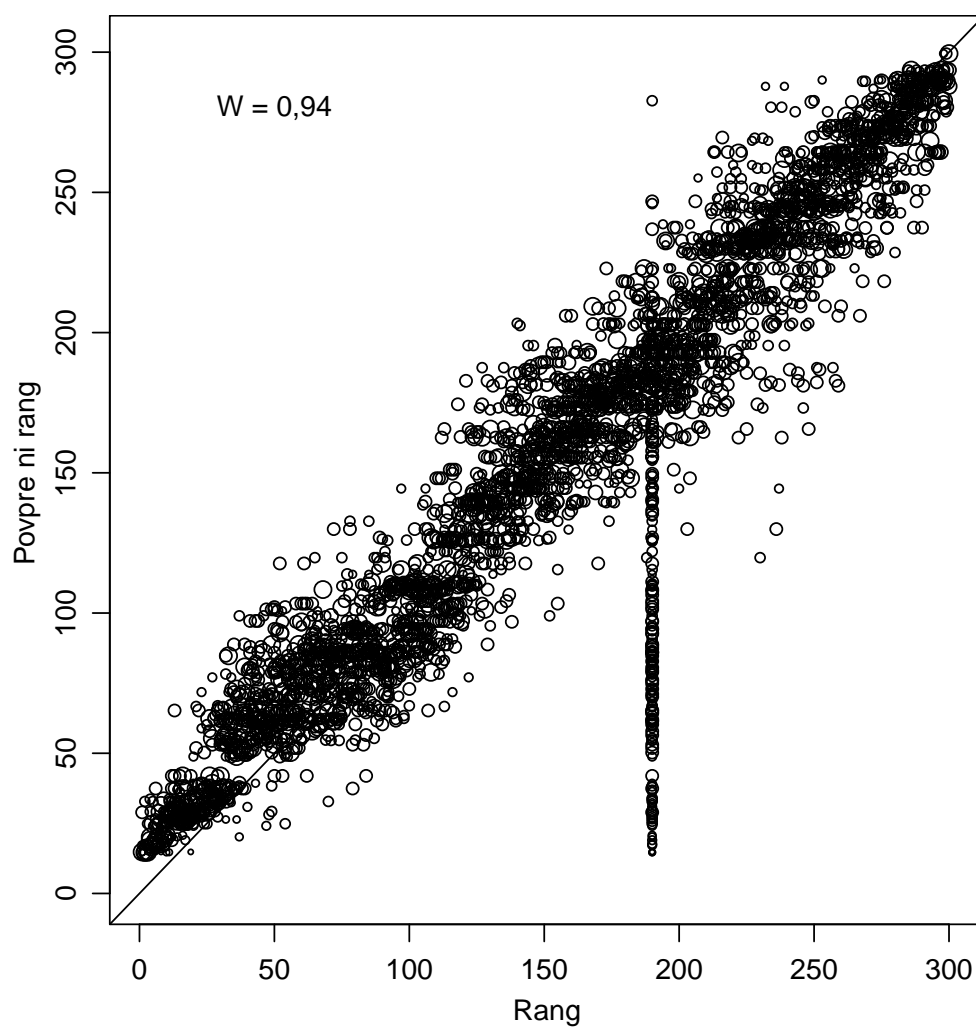


Slika 4.1.5: Prikaz konkordance 16 mer za vrednotenje uvrščanja s Kendallovim koeficientom konkordance ( $W$ ) in s prikazom rangov na vzporednih oseh za modele uvrščanja zgrajene z metodami 1-NN, 3-NN in 5-NN ter ovrednotene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi. Modeli so za vsako od metod uvrščanja razporejeni od najboljšega (zgoraj) do najslabšega (spodaj) glede na mero A-povprečje.





Slika 4.1.6: Grafični prikaz Kendallovih korelacijskih koeficientov ( $\tau$ ), izračunanih za vse pare 16 obravnavanih mer za vrednotenje uvrščanja, ki so bile izračunane za 188 modelov, ovrednotenih s 100 krat ponovljenim prečnim preverjanjem s pregibanjem (razdelek 4.1.1), in 112 modelov ovrednotenih s 500 krat ponovljeno razdelitvijo na testno in učno množico (razdelek 4.1.2).

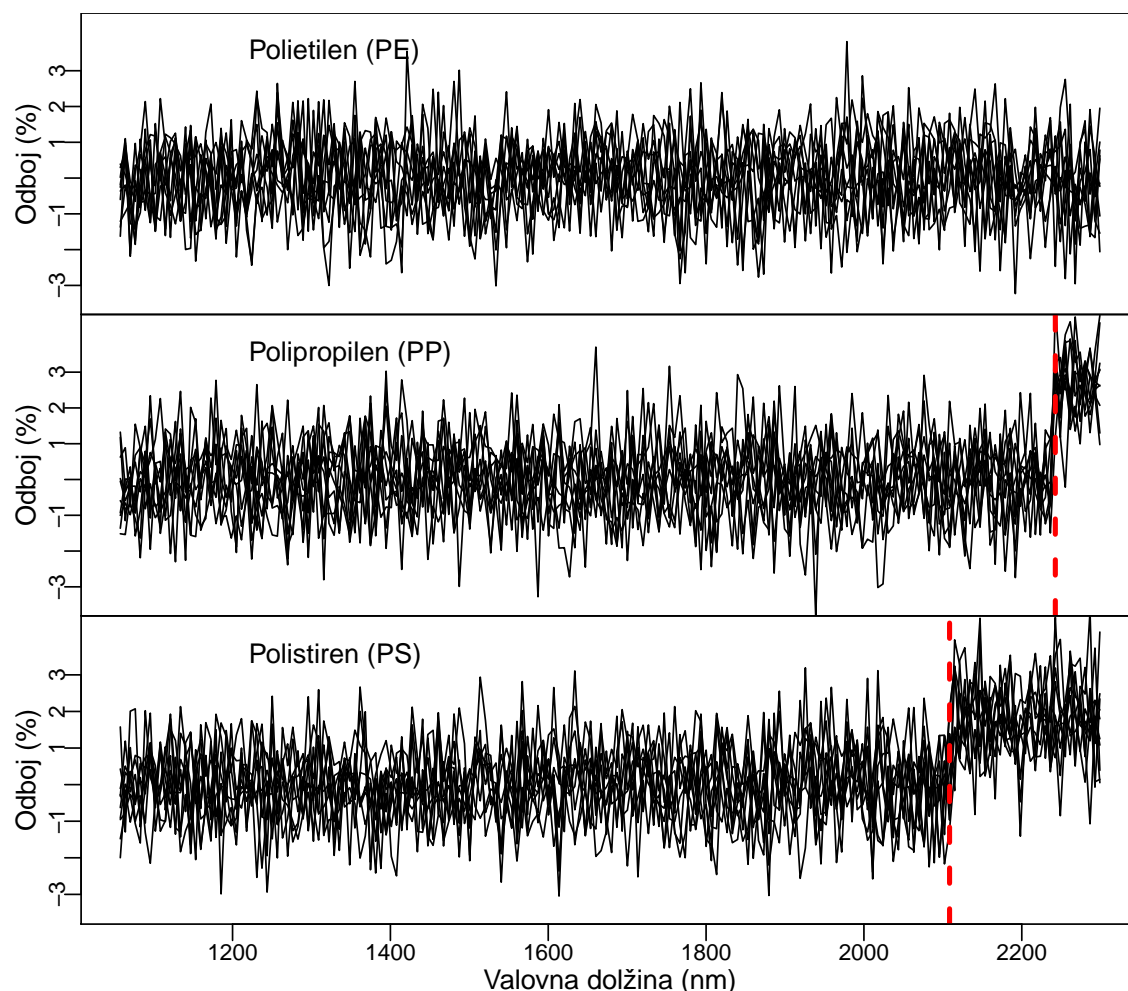


Slika 4.1.7: Konkordančni mehurčni diagram z izračunanim Kendallovim koeficientom konkordance ( $W$ ) za 16 obravnavanih mer za vrednotenje uvrščanja, ki so bile izračunane za 188 modelov, ovrednotenih s 100 krat ponovljenim prečnim preverjanjem s pregibanjem (razdelek 4.1.1), in 112 modelov, ovrednotenih s 500 krat ponovljeno razdelitvijo na testno in učno množico (razdelek 4.1.2).

## 4.2 Generirani spektri

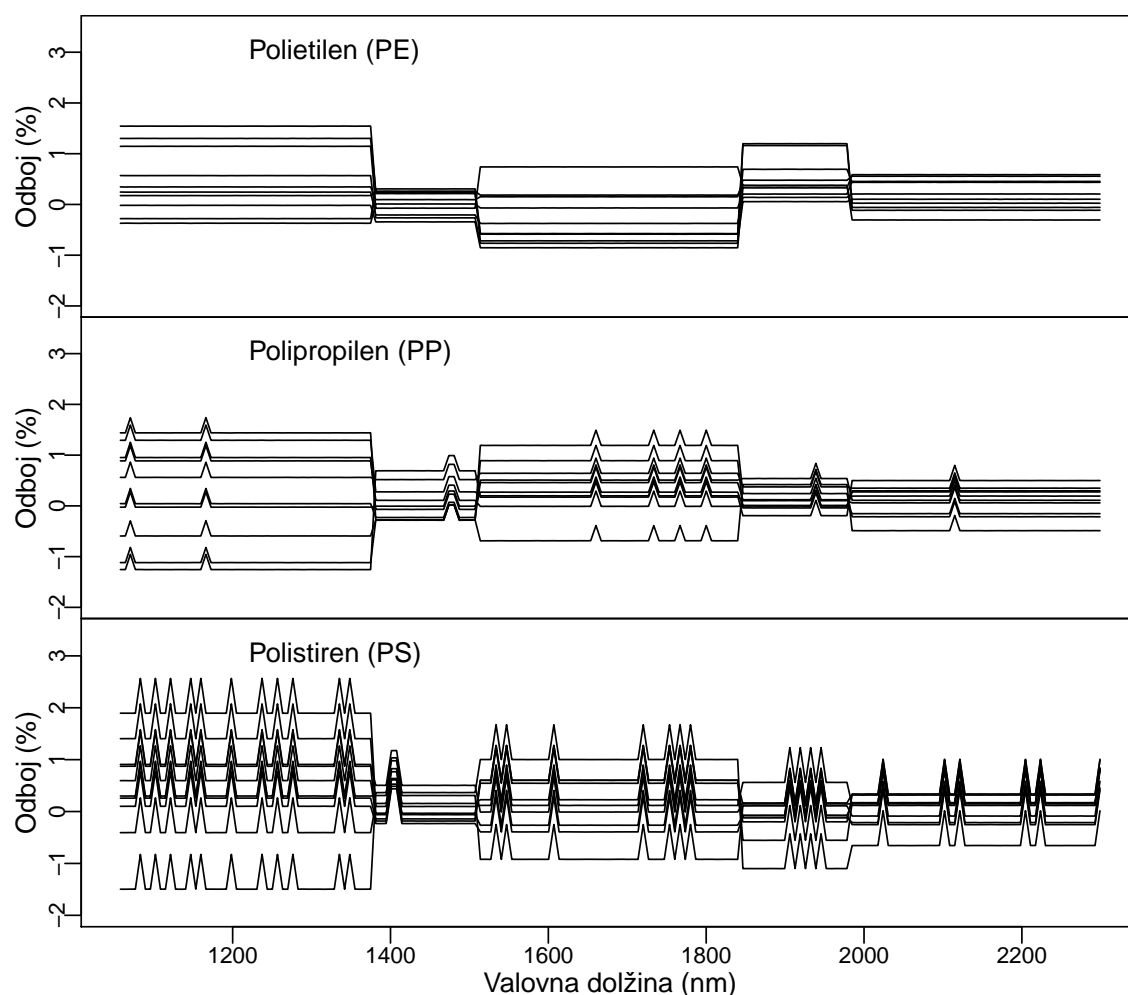
V tem poglavju so predstavljeni podatki, ki smo jih generirali na štiri različne načine, kot je opisano v razdelku 3.4. Pri generiranju podatkov smo se osredotočili le na tri razrede, s katerimi smo se želeli čim bolj približati dejanskim NIR meritvam polimerov PE, PP in PS, ki so prikazane na Sliki 2.1.1.

Pri prvem načinu (IND) smo generirali podatke neodvisno iz normalne porazdelitve  $N(\mu, \sigma)$ . Razlike med razredi smo generirali s pomočjo razlik v parametru  $\mu$ , ki je bil v razredu PP pri 10 spremenljivkah, v razredu PS pa pri 30 spremenljivkah različno izražen kot v razredu PE. Točne vrednosti parametra  $\mu$  pri različno izraženih spremenljivkah so predstavljene v Tabeli 4.2.1. Natančen opis generiranja podatkov po metodi IND je v razdelku 3.4.1, deset generiranih enot iz vsakega razreda je prikazanih na Sliki 4.2.1.



Slika 4.2.1: Generirani spektri po metodi IND. Za vsak razred je prikazanih 10 enot. Zadnjih 10 spremenljivk pri polipropilenu (PP) in zadnjih 30 spremenljivk pri polistirenu (PS) je različno izraženih (spremenljivke za črtkano črto na sliki).

Pri drugem načinu smo povezanost med spremenljivkami generirali s pomočjo multivariatne normalne porazdelitve z bločno kovariančno matriko (MVNblock), razlike med skupinami smo generirali z različnimi vektorji povprečij. Vektor povprečij se je za razred PP pri 10 komponentah za razred PS pa pri 30 komponentah razlikoval od vektorja povprečij za razred PE. Točne vrednosti različno izraženih komponent so predstavljene v Tabeli 4.2.1. Natančen postopek generiranja MVNblock podatkov je opisan v razdelku 3.4.2, 10 enot iz vsakega razreda je prikazanih na Sliki 4.2.2.



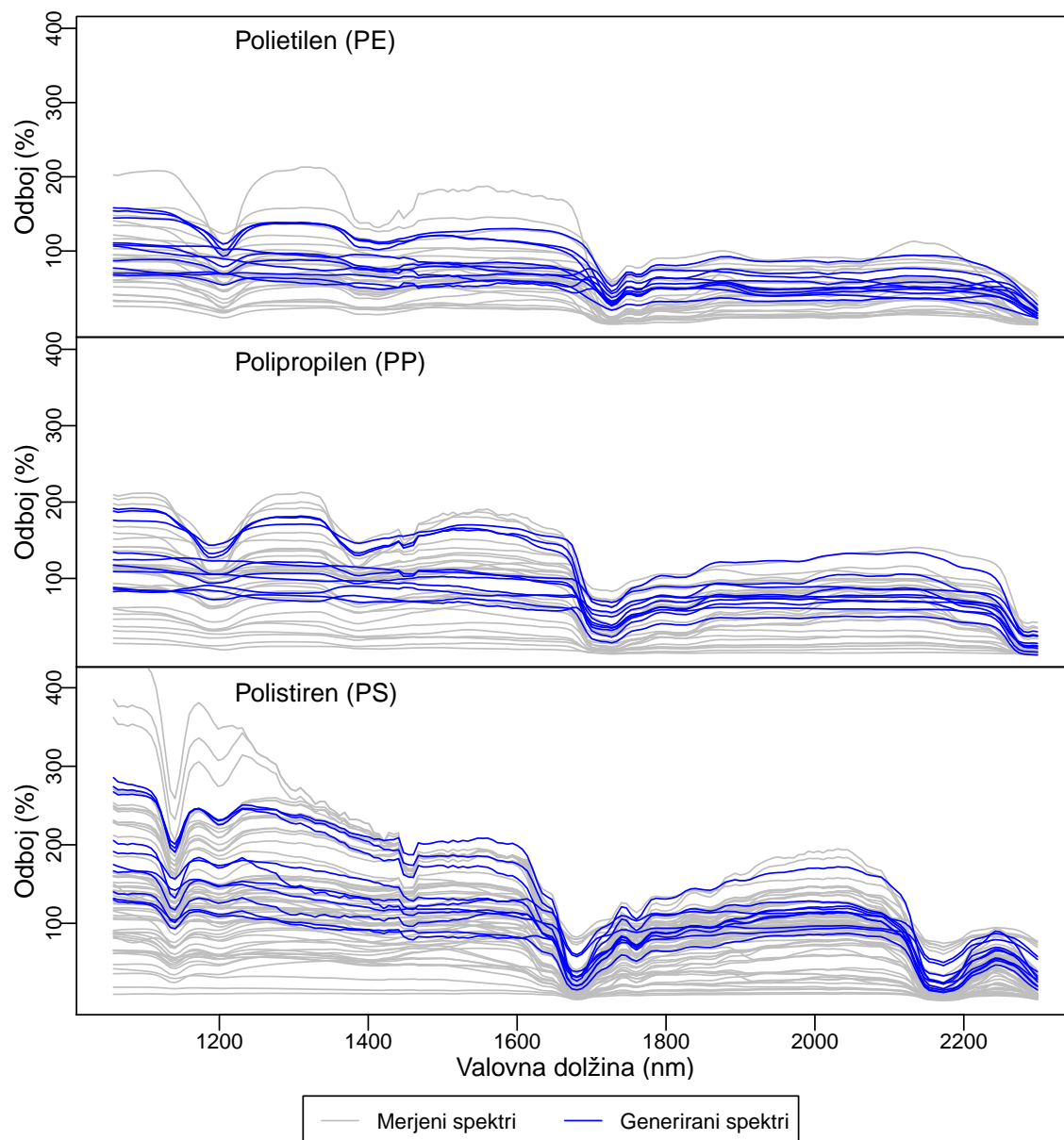
Slika 4.2.2: Generirani spektri po metodi MVNblock. Za vsak razred je prikazanih 10 enot.

	IND		MVNblock	
	PP	PS	PP	PS
LDA	2,69	1,72	0,0034	0,0024
CART	1,24	1,98	0,30	0,67
SVM	2,05	1,11	0,085	0,059

Tabela 4.2.1: Vrednosti parametrov  $\mu$  in  $M$  za zadnjih 10 spremenljivk pri razredu PP in 30 naključno izbranih spremenljivk pri razredu PS pri generiranju neodvisnih podatkov (IND) in pri generiranju koreliranih podatkov z bločno kovariančno matriko (MVNblock).

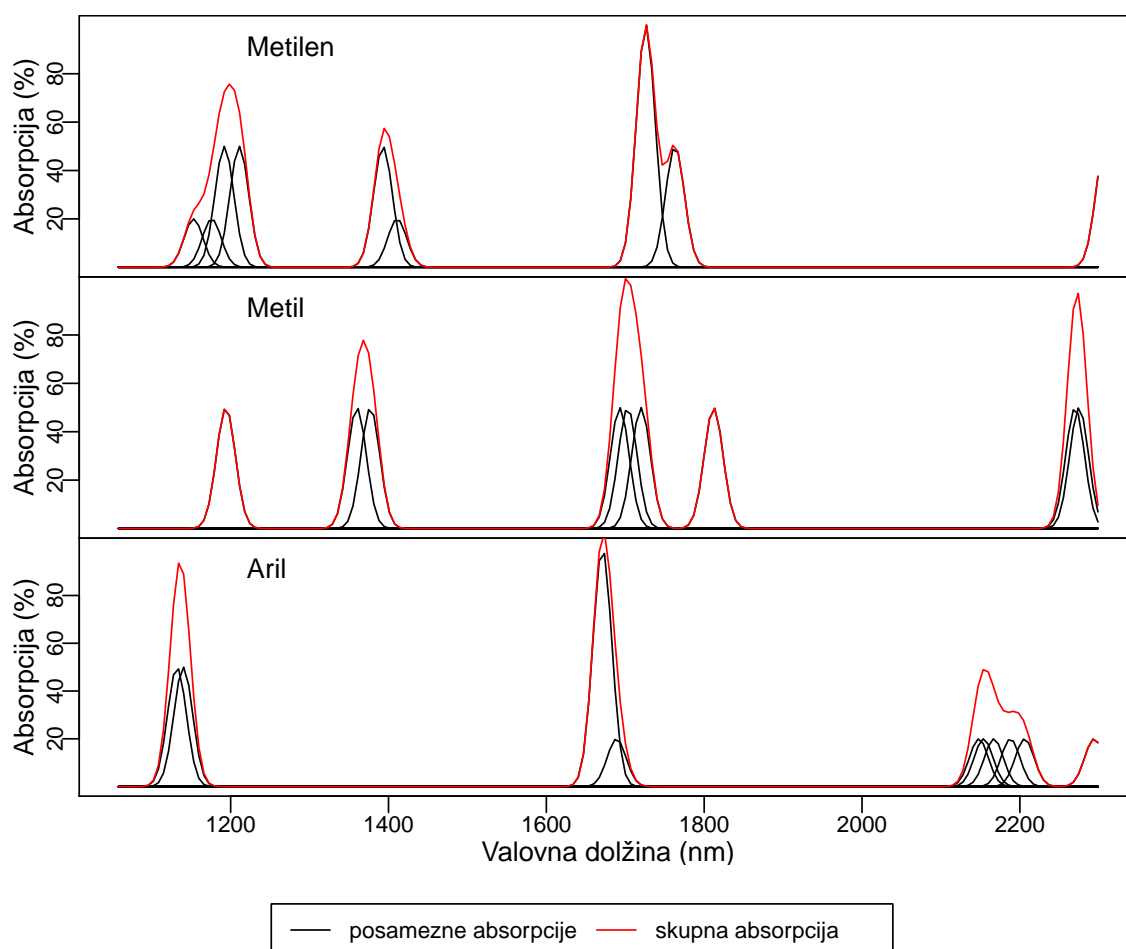
Tudi pri tretjem načinu generiranja podatkov (MVNorig) smo uporabili multivariatno normalno porazdelitev  $MVN(M, \Sigma)$ , kjer smo parametra  $M$  in  $\Sigma$  ocenili iz realnih podatkov

polimerov kot je opisano v razdelku 3.4.3. Ti spektri se bolje prilegajo realnim podatkom kot spektri generirani na prva dva načina, zato je na Sliki 4.2.3 prikazanih po 10 enot iz vsakega razreda generiranih z metodo MVNorig skupaj z realnimi podatki.



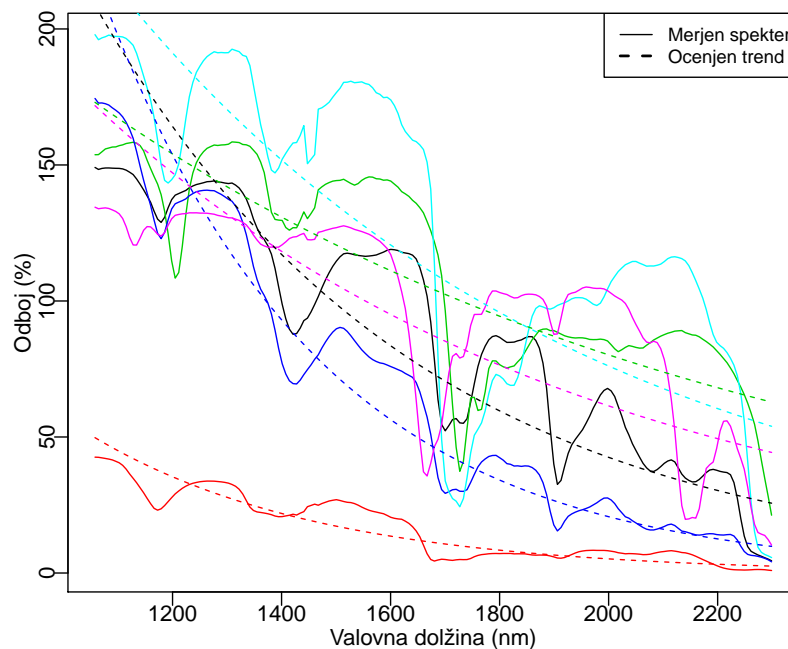
Slika 4.2.3: Generirani spektri po metodi MVNorig (za vsak razred je prikazanih 10 enot) in merjeni spektri polimerov PE, PP in PS iz množice realnih podatkov.

Pri četrtem načinu generiranja podatkov (ABS) smo upoštevali informacijo o kemijski strukturi polimerov PE, PP in PS ter absorpciji funkcionalnih skupin v NIR območju, kot je opisano v razdelku 3.4.4. V ta namen smo iz literature dobili informacijo o absorpcijskih mestih in njihovi moči za posamezne funkcionalne skupine, ki smo jih nato s seštevanjem združili v skupno absorpcijsko krivuljo funkcionalne skupine (razdelek 3.4.4.3). Na Sliki 4.2.4 so prikazane absorpcije metilne, metilenske in arilne funkcionalne skupine. Vidimo lahko, da sta si krivulji metilenske in metilne skupine veliko bolj podobni kot krivulji metilenske in arilne ali metilne in arilne funkcionalne skupine, kar ustreza podobnosti v kemijski zgradbi med opazovanimi funkcionalnimi skupinami.



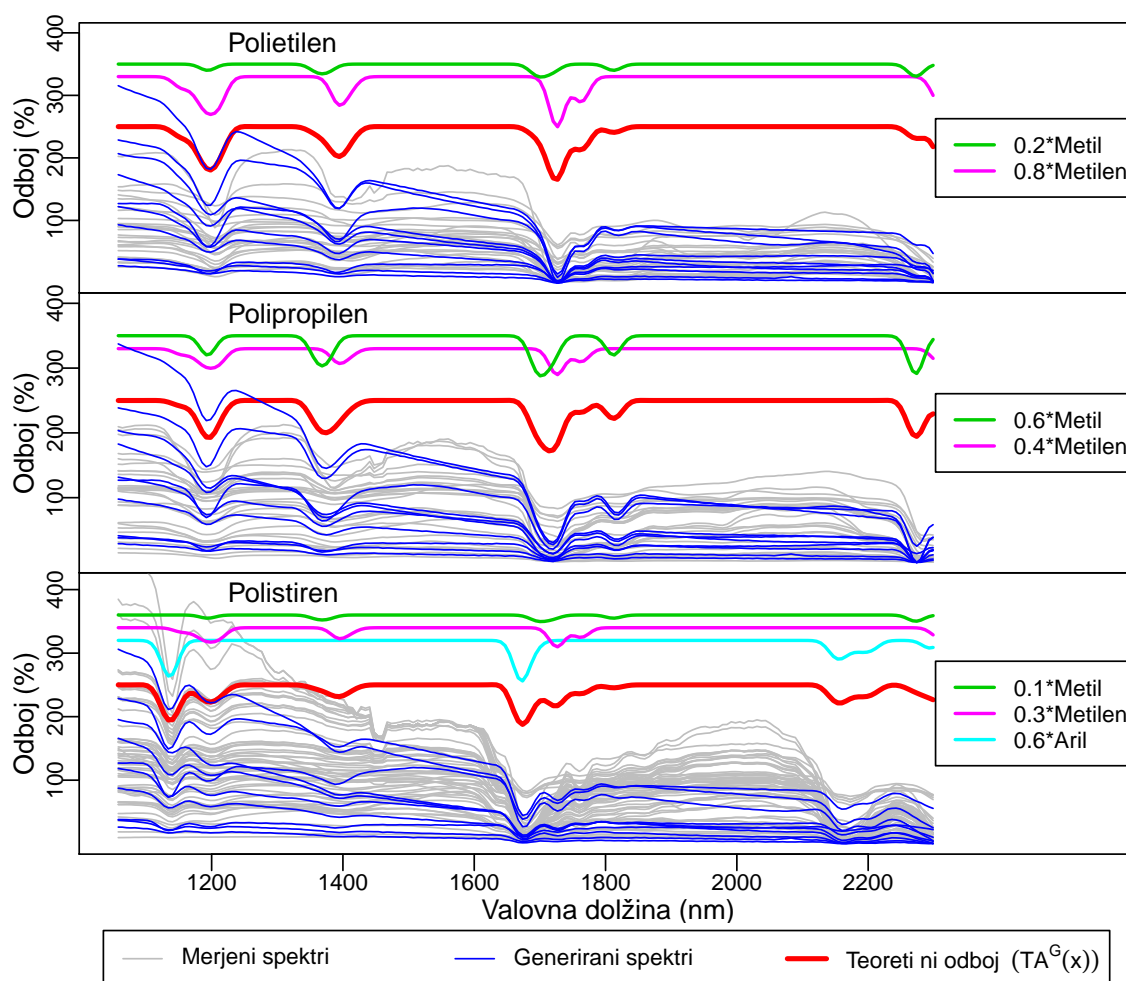
Slika 4.2.4: Absorpcijska mesta in skupna absorpcijska krivulja posamezne funkcionalne skupine.

V 534 spektrih odboja v NIR območju smo opazili padajoči trend, ki smo ga vključili v metodo generiranja podatkov po metodi ABS. Predpostavili smo, da ima trend obliko eksponentne krivulje. Da bi lažje opazovali prilaganje eksponentne trendne krivulje dejanskim meritvam, je na Sliki 4.2.5 prikazanih šest naključno izbranih spektrov iz zbirke 534 polimernih vzorcev z ocenjenim eksponentnim trendom. Na skrajnih robovih opazovanega območja ležijo trendne krivulje v večini primerov nad pripadajočimi merjenimi spektri, medtem ko se v širši sredini opazovanega območja sorazmerno dobro prilagajo opazovanim vrednostim.



Slika 4.2.5: Šest naključno izbranih enot iz zbirke 534 spektrov polimerov z ocenjenim eksponentnim trendom.

Na Sliki 4.2.6 so ob enotah generiranih spektrov po metodi ABS in realnih podatkov prikazane še z ustrezno utežjo pomnožene skupne teoretične absorpcije funkcionalnih skupin in skupna absorpcija posameznega polimera, ki smo jih pri postopku generiranja podatkov potrebovali. Natančen opis teh teoretičnih absorpcij je v razdelku 3.4.4.



Slika 4.2.6: Realni podatki in podatki, generirani na podlagi teoretičnih absorpcij funkcionalnih skupin. Na sliki so prikazane tudi z ustrežno utežjo pomnožene skupne teoretične absorpcije metilne, metilenske in arilne funkcionalne skupine, ki smo jim spremenili predznak, da jih lažje primerjamo s spektri odboja. Skupne teoretične absorpcije funkcionalnih skupin so prikazane zamaknjeno na  $y$ -osi zaradi boljše preglednosti. Podobno je zamaknjeno na  $y$ -osi in s spremenjenim predznakom prikazana skupna absorpcija posameznega materiala ( $TA^G(x)$ ).



## 4.3 Rezultati simulacij

Razdelek Rezultati simulacij je razdeljen na štiri dele. V prvem 4.3.1 so prikazani rezultati uvrščanja z metodo SVM ob uporabi linearne ali radialne jedrne funkcije ter metod za prilagoditev uvrščanja na velikost razredov z namenom poiskati najboljšo kombinacijo pri uvrščanju z metodo SVM. V drugem razdelku 4.3.2 so prikazani rezultati simulacij uvrščanja v dva razreda, v tretjem 4.3.3 rezultati simulacij uvrščanja v tri razrede in v četrtem 4.3.4 rezultati simulacij uvrščanja v 45 razredov. Pri uvrščanju v dva in tri razrede smo v simulacijah uporabili vse štiri metode generiranja podatkov (IND, MVNblock, MVNreal in ABS), opisane v poglavju 3.4 in realne podatke polimerov. Tako realni podatki kot generirani podatki so bili sestavljeni iz skupin PE, PP in PS z velikostjo  $n_{PE} = 26$ ,  $n_{PP} = 27$  in  $n_{PS} = 59$ , kjer so bile med skupinama PE in PP manjše razlike kot med skupinama PE in PS ali PP in PS. Pri simulacijah uvrščanja v 45 razredov smo uporabili le eno metodo za umetno generiranje NIRS podatkov, in sicer MVNorig. Kako so bile določene razlike med razredi, da so primerljive z realnimi podatki, je opisano pri vsaki metodi generiranja podatkov posebej v razdelku 3.4.

Na slikah, v katerih so prikazani rezultati uvrščanja v dva razreda (Slike 4.3.1–4.3.10), so prikazane povprečne vrednosti mere G, izračunane iz 1000 ponovitev simulacij pri uvrščanju z metodami LDA, CART in L-SVM. Vrednosti mere G v vsaki ponovitvi simulacije so bile pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi. Simulirali smo uvrščanje v dva razreda, med katerima ni bilo razlik (PS-PS), so bile majhne razlike (PE-PP) ali so bile velike razlike (PE-PS). Na vsaki sliki so prikazani rezultati uvrščanja ob različnih stopnjah neravnotežja med razredoma, ki je določeno z deležem enot prvega razreda  $k_1$ . Število enot, ki smo jih v vsakem razredu generirali, je bilo takšno kot v realnih podatkih:  $n_{PE} = 26$ ,  $n_{PP} = 27$  in  $n_{PS} = 59$ .

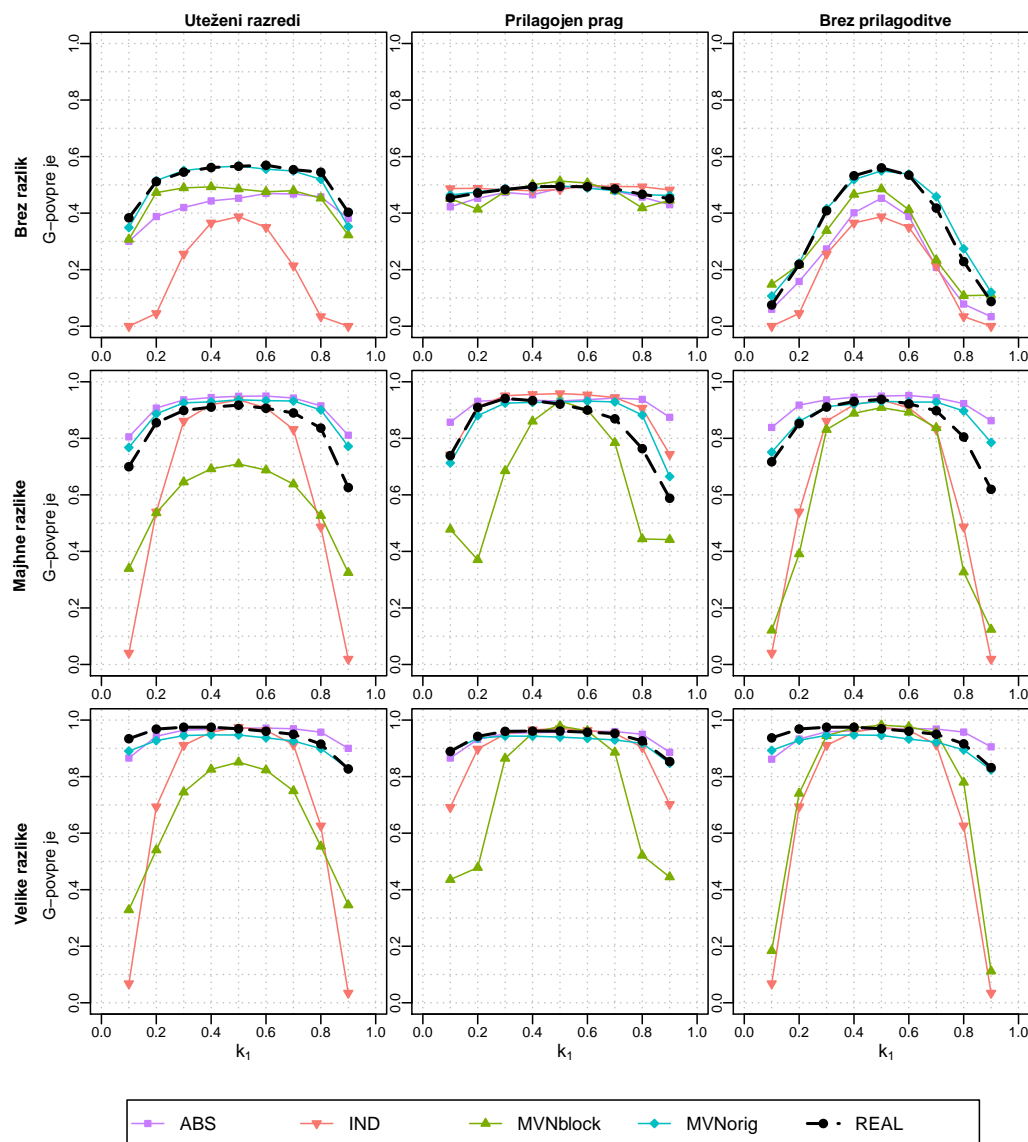
Vrednosti mere G so visoke, če sta napovedni točnosti obeh razredov visoki. Nizka vrednost mere G lahko pove, da je nizka le napovedna točnost enega razreda, ali pa sta nizki napovedni točnosti obeh razredov. Vpliv neravnotežja se kaže tako, da so napovedne točnosti večjega razreda visoke, napovedne točnosti manjšega razreda pa nizke, kar povzroči nizko vrednost mere G. Na grafih v tem razdelku, kjer imamo vrednosti mere G, prikazane pri različnih stopnjah neravnotežja, lahko opazimo značilno obliko, ki jo dobimo, če te točke med seboj povežemo. Krivulja, ki povezuje vrednosti mere G pri različnih stopnjah neravnotežja je običajno najvišja v uravnoteženem primeru ( $k_1 = 0,5$ ), nato pa z večanjem neravnotežja simetrično pada (ko gre  $k_1$  od 0,5 do 0,1 oz. od 0,5 proti 0,9). Bolj strm padec krivulje, ki povezuje vrednosti mere G pri različnih stopnjah neravnotežja, zato nakazuje večji vpliv neravnotežja.

Na slikah, v katerih so prikazani rezultati uvrščanja v tri razrede (Slike 4.3.11–4.3.15 in v dodatku D.1–D.9), so prikazane povprečne napovedne točnosti po razredih, izračunane iz 1000 ponovitev simulacij uvrščanja v tri razrede z metodami LDA, CART in L-SVM. Povprečne napovedne točnosti po razredih so bile pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi. Simulirali smo uvrščanje v tri razrede, kjer so bili vsi trije razredi enaki (PS-PS-PS), kjer sta bila dva razreda enaka in eden različen (PP-PP-PE, PS-PS-PE) in kjer so bili vsi trije razredi različni (PS-PP-PE). Pri tem so bile razlike med razredoma PP in PE manjše kot med razredoma PS in PE ali PS in PP. Pri simulacijah smo spreminjali velikost razredov, in s tem stopnjo neravnotežja.

V primerih, kjer so bile nastavitve simulacij nekoliko drugačne od zgoraj opisanih, je to

napisano naknadno ob rezultatih. Potek simulacij in pri simulacijah uporabljene metode so natančno opisane v razdelku 3.5.

#### 4.3.1 SVM z linearno in radialno jedrno funkcijo (Sliki 4.3.1 in 4.3.2)

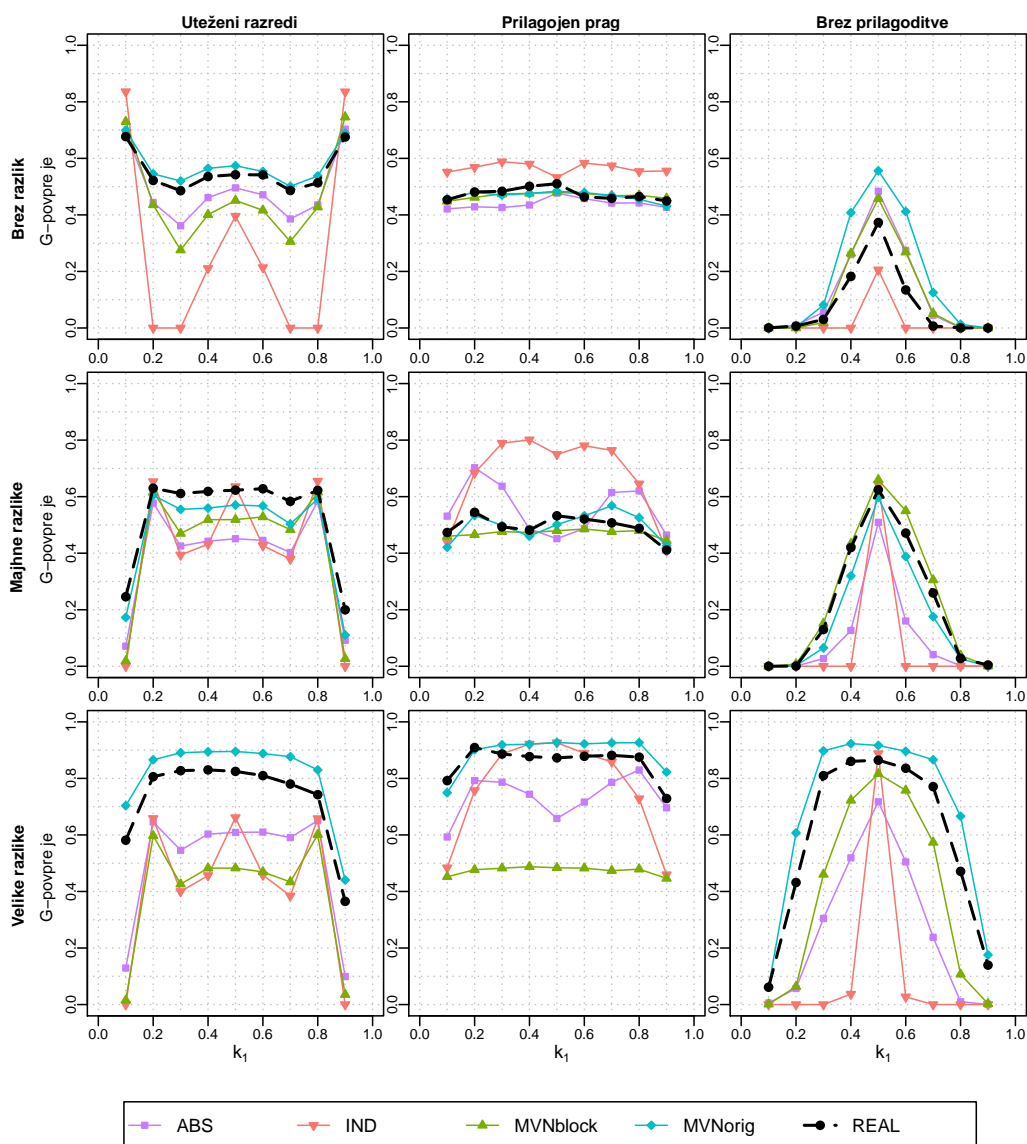


Slika 4.3.1: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodo SVM, pri kateri smo uporabili linearno jedrno funkcijo, vpliv neravnotežja pa smo zmanjšali z uporabo uteži in spremembo praga za uvrščanje.

S simulacijami uvrščanja v dva razreda smo najprej želeli proučiti vpliv jedrne funkcije (radialna ali linearna), uteževanja razredov (z možnostjo *class.weights* v funkciji *svm*) in spremembe praga za uvrščanje glede na velikost razredov pri uvrščanju z metodo SVM (Slika 4.3.1 in Slika 4.3.2).

Rezultati uvrščanja so bili boljši in vpliv neravnotežja manjši, ob uporabi linearne jedrne

funkcije v primerjavi z rezultati, pridobljenimi z radialno jedrno funkcijo tako v primeru brez prilagoditve na velikost razredov kot v primeru uteževanja in prilagoditve praga za uvrščanje. Obe metodi prilagoditve na velikost razredov sta zmanjšali vpliv neravnotežja, pri čemer je bila metoda prilagoditve praga za uvrščanje uspešnejša od uteževanja. V vseh obravnavanih primerih uvrščanja z metodo SVM je vpliv neravnotežja na uvrščanje najbolj izrazit za nekorelirane podatke (IND) in podatke generirane z bločno kovariančno matriko (MVNblock). Ker se je klasifikator, zgrajen z metodo SVM, linearno jedrno funkcijo in prilagojenim pragom za uvrščanje izkazal za najuspešnejšega med SVM klasifikatorji, so v nadaljevanju prikazani rezultati simulacij uvrščanja z metodo SVM s temi nastavitvami (L-SVM).



Slika 4.3.2: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodo SVM, pri kateri smo uporabili radialno jedrno funkcijo, vpliv neravnotežja pa smo zmanjšali z uporabo uteži in s spremembo praga za uvrščanje.

### 4.3.2 Uvrščanje v dva razreda

S simulacijami uvrščanja v dva razreda smo želeli proučiti vpliv predobdelave (SNV in 1. odvod), zmanjšanja dimenzije podatkov (izbor 50 spremenljivk z največjo varianco, izbor 50 spremenljivk z največjo F-statistiko in metoda PCA) in metode MDS na uvrščanje NIRS podatkov. Ob tem smo želeli proučiti, kako dobro se rezultati umetno generiranih podatkov približajo realnim.

#### 4.3.2.1 Brez predobdelave in brez zmanjšanja dimenzije podatkov (Sliki 4.3.3 in 4.3.4)

Ko ni bila uporabljena nobena metoda predobdelave ali metoda zmanjšanja dimenzije podatkov (Slika 4.3.3), lahko pri realnih podatkih opazimo, da so se vrednosti mere  $G$  z večanjem razlik med razredi povišale pri vseh treh metodah uvrščanja. Pri tem so razlike med razredi najmanj vplivale na metodo CART, saj so se vrednosti mere  $G$  pri CART z razlikami med razredi manj povišale kot pri LDA in L-SVM. Pri realnih podatkih lahko v primeru brez razlik med razredoma opazimo, da je bila na neravnotežje najmanj občutljiva metoda L-SVM, najbolj občutljiva pa metoda CART. Z večanjem razlik med razredi se je vpliv neravnotežja pri metodah LDA in CART zmanjševal, pri metodi L-SVM pa je bil vpliv neravnotežja največji pri majhnih razlikah med razredi.

Pri IND podatkih opazimo večji vpliv neravnotežja kot pri realnih podatkih – vrednosti mere  $G$  v uravnoteženem primeru ( $k_1 = 0,5$ ) se z večanjem razlik med razredi zvišujejo, medtem, ko vrednosti mere  $G$  v močno neuravnoteženem primeru ( $k_1 = 0,1$ ) ostajajo nizke. To je najbolj izrazito pri uvrščanju z LDA. Z IND podatki bi torej močno precenili vpliv neravnotežja v primerjavi z realnimi podatki (Slika 4.3.3).

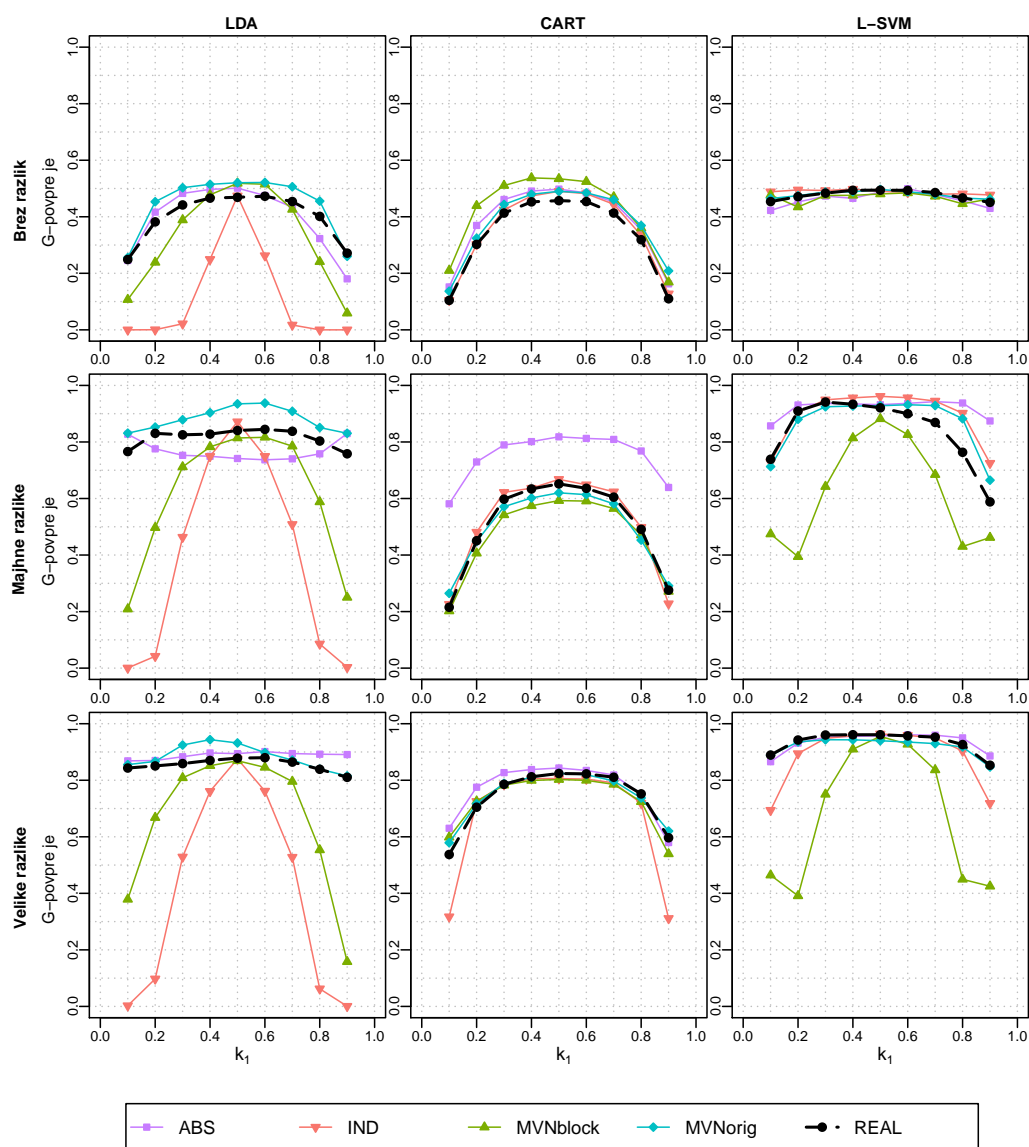
Podobno kot z IND podatki bi tudi z MVNblock podatki precenili vpliv neravnotežja realnih podatkov predvsem pri uvrščanju z LDA in L-SVM (Slika 4.3.3).

Rezultati uvrščanja ABS podatkov se dobro prilegajo realnim podatkom v primerih brez razlik in z velikimi razlikami med razredoma. V primeru manjših razlik med razredoma pa kažejo manjši vpliv neravnotežja kot realni podatki (Slika 4.3.3).

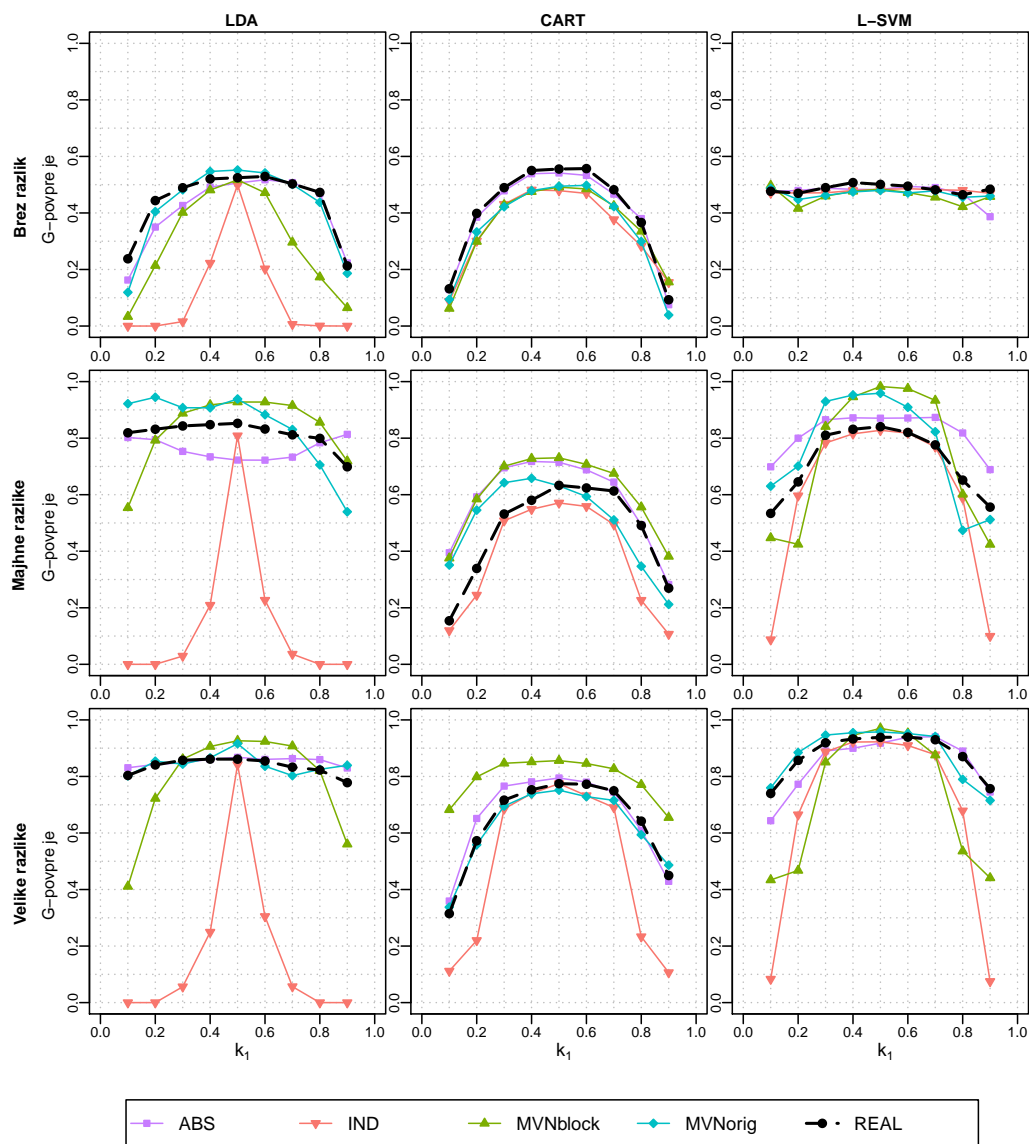
Rezultati uvrščanja MVNorig se v vseh obravnavanih primerih uvrščanja na Sliki 4.3.3 najbolj prilegajo realnim podatkom.

Podobnost med rezultati uvrščanja MVNorig podatkov in realnih podatkov, ki smo jo opazili na Sliki 4.3.3, bi lahko bila posledica odvisnosti MVNorig podatkov od realnih podatkov. Za ocenitev parametrov MVN porazdelitve, ki smo jih uporabili pri generiranju MVNorig podatkov, smo namreč uporabili realne podatke, ki so bili hkrati uporabljeni za simulacijo uvrščanja realnih podatkov (REAL). Zato smo v nadaljevanju 30 enot realnih podatkov (10 naključno izbranih enot iz vsakega razreda) uporabili za ocenitev parametrov MVN porazdelitve pri generiranju MVNorig podatkov. Teh 30 enot nato nismo vključili v simulacijo uvrščanja realnih podatkov v dva razreda. Število enot realnih podatkov v vsakem razredu, ki so bile vključene v simulacijsko študijo, je tako bilo  $n_{PE} = 16$ ,  $n_{PP} = 17$  in  $n_{PS} = 49$ . Da bi se realnemu stanju čim bolj približali, so bile takšne tudi velikosti razredov umetno generiranih podatkov (Slika 4.3.4). Opazimo lahko, da se tudi v tem primeru rezultati uvrščanja MVNorig podatkov dobro približajo realnim podatkom, kar kaže na to, da podobnost med rezultati MVNorig in realnimi podatki, ki smo jo opazili

na Sliki 4.3.3, ni bila le posledica neposredne odvisnosti MVNorig podatkov od realnih. Večja asimetrija in večji vpliv neravnotežja, ki ga lahko na splošno opazimo pri vseh vrstah podatkov, če Sliko 4.3.4 primerjamo s Sliko 4.3.3 lahko najverjetneje pripišemo manjšemu številu enot v razredih.



Slika 4.3.3: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM brez uporabe predobdelave, metod za zmanjšanje dimenzije podatkov ali metode MDS. Parametri multivariatne normalne porazdelitve pri generiranju MVNorig podatkov so ocenjeni iz podatkov REAL.



Slika 4.3.4: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM brez uporabe predobdelave, metod za zmanjšanje dimenzije podatkov ali metode MDS. Parametri multivariatne normalne porazdelitve pri generiranju MVNorig podatkov so neodvisni od podatkov REAL, zato je bilo število generiranih enot v posameznem razredu manjše ( $n_{PE} = 16$ ,  $n_{PP} = 17$  in  $n_{PS} = 49$ ).

#### 4.3.2.2 Predobdelave spektrov (Slika 4.3.5 in Slika 4.3.6)

Pri uporabi predobdelave SNV (Slika 4.3.5) na realnih podatkih opazimo, da so se vrednosti mere  $G$  zvišale, vpliv neravnotežja pa se je znižal v primerjavi z rezultati brez uporabe predobdelave (Slika 4.3.3). To opazimo pri vseh treh opazovanih metodah uvrščanja, najbolj izrazito pri metodi CART.

Metoda SNV je imela na IND podatke večinoma prav nasproten vpliv kot na realne podatke: vrednosti mere  $G$  pri uvrščanju IND podatkov so večinoma nižje, vpliv neravnotežja pa višji ob uporabi SNV kot brez uporabe predobdelave (Slika 4.3.3).

Metoda SNV je podobno kot pri uvrščanju realnih podatkov tudi pri uvrščanju MVNblock podatkov znižala vpliv neravnotežja v primerjavi z uvrščanjem brez uporabe predobdelave (Slika 4.3.3), kar je bilo bolj izrazito v primeru velikih razlik med razredoma. Tudi ob uporabi SNV pa bi z MVNblock podatki v primeru majhnih razlik med razredoma močno precenili vpliv neravnotežja pri realnih podatkih, še posebej pri uvrščanju z LDA in L-SVM.

Na podatke ABS je metoda SNV vplivala podobno kot na realne podatke: vrednosti mere  $G$  so se povečale, vpliv neravnotežja se je zmanjšal v primerjavi z rezultati brez uporabe SNV. Po uporabi metode SNV se rezultati uvrščanja ABS podatkov bolje prilegajo realnim podatkom kot brez uporabe predobdelave (Slika 4.3.3). Največje odstopanje rezultatov ABS od realnih podatkov podobno kot v primeru brez uporabe predobdelave opazimo v primeru uvrščanja med razredoma z majhnimi razlikami (Slika 4.3.5).

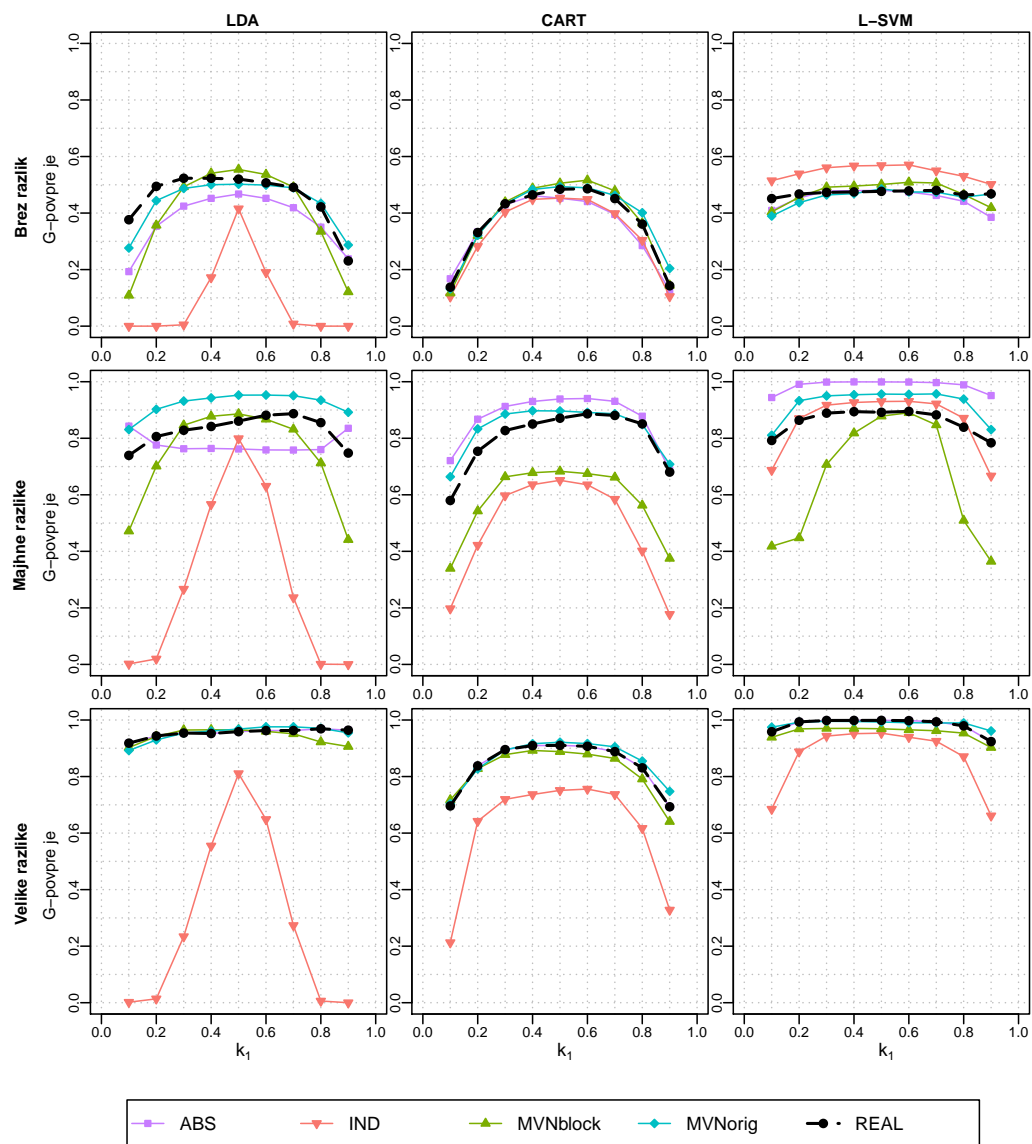
MVNorig podatki se tudi pri uporabi predobdelave SNV najboljše prilegajo realnim podatkom, kar pomeni, da uporaba SNV podobno vpliva na MVNorig podatke kot na realne podatke (Slika 4.3.5).

Rezultati realnih podatkov ob uporabi predobdelave 1. odvod podobno kot predobdelava SNV izboljšajo rezultate uvrščanja brez uporabe predobdelave: opazimo lahko višje vrednosti mere  $G$  in zmanjšan vpliv neravnotežja pri uvrščanju z vsemi tremi metodami uvrščanja, ko so med razredi razlike. Še posebej izrazito je pri uvrščanju z metodo CART (Slika 4.3.6).

Na uvrščanje IND podatkov je predobdelava 1. odvod drugače vplivala kot na realne podatke: vpliv neravnotežja se je sicer ob uporabi 1. odvoda v večini primerov nekoliko zmanjšal v primerjavi z rezultati brez uporabe predobdelave (Slika 4.3.3), zmanjšale so se tudi vrednosti mere  $G$  pri vseh stopnjah neravnotežja. Posledično se IND podatki ob uporabi 1. odvoda še slabše prilegajo realnim podatkom kot brez uporabe predobdelave (Slika 4.3.6).

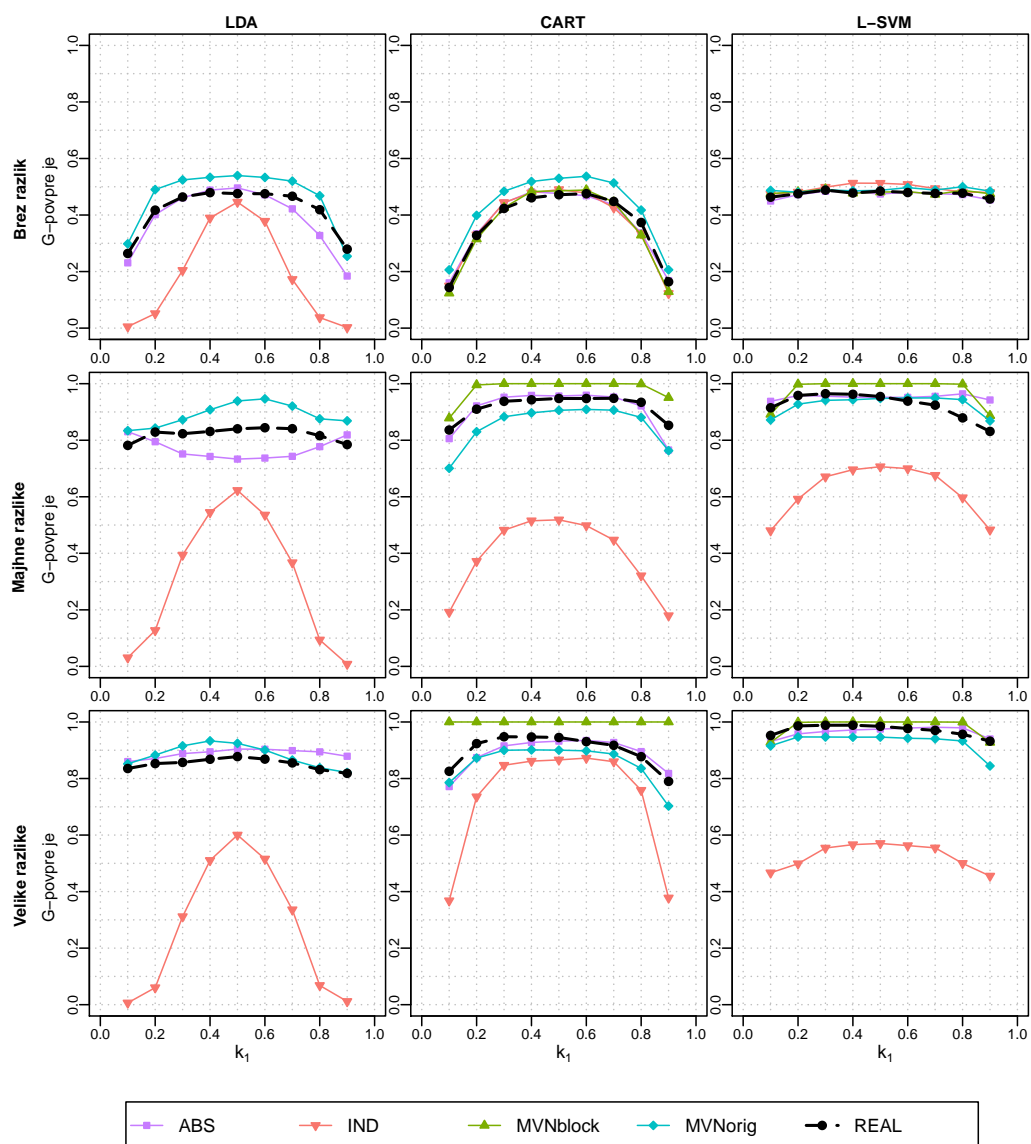
Vrednosti mere  $G$  so pri uvrščanju MVNblock podatkov z metodo CART ob uporabi 1. odvoda, ko so med razredi razlike, zelo visoke (skoraj 1 pri vseh stopnjah neravnotežja), zaradi česar vpliva neravnotežja skoraj ni opaziti. V tem primeru imajo rezultati MVNblock podatkov veliko višje vrednosti mere  $G$  kot realni podatki. To pa je prav nasprotno opažanjem pri rezultatih brez uporabe predobdelave (Slika 4.3.3), kjer so imeli MVNblock podatki pri uvrščanju s CART nekoliko nižje vrednosti mere  $G$  in podoben vpliv neravnotežja kot realni podatki (Slika 4.3.6). Pri uvrščanju z L-SVM ob uporabi 1. odvoda se vpliv neravnotežja pri MVNblock podatkih močno zmanjša v primerjavi z rezultati brez uporabe predobdelave (Slika 4.3.3), zato ob uporabi 1. odvoda MVNblock podatki kažejo podoben vpliv neravnotežja in podobne vrednosti mere  $G$  kot realni podatki.

Uporaba 1. odvoda na uvrščanje podatkov ABS in MVNorig podobno vpliva kot na realne podatke (Slika 4.3.6).



Slika 4.3.5: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM ob uporabi predobdelave SNV.





Slika 4.3.6: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM. Na enotah je bil predhodno izračunan 1. odvod s pomočjo Savitsky-Golay filtra z oknom 7 in 1. stopnjo polinoma. Na podatkih MVNblock modela uvrščanja z metodo LDA ni bilo mogoče zgraditi zaradi kolinearnosti med spremenljivkami.

#### 4.3.2.3 Zmanjšanje dimenzije podatkov (Slike 4.3.7, 4.3.8 in 4.3.9)

Pri uvrščanju realnih podatkov ob uporabi izbora spremenljivk glede na največjo varianco (Slika 4.3.7) opazimo nižje vrednosti mere  $G$ , vpliv neravnotežja pa ostane približno enak kot v primeru brez zmanjšanja dimenzije podatkov (Slika 4.3.3). Izjema pri tem je uvrščanje brez razlik med razredi z metodo LDA, kjer se je vpliv neravnotežja zmanjšal, in uvrščanje brez razlik med razredi z metodo CART, kjer se vrednosti mere  $G$  v primerjavi z rezultati brez zmanjšanja dimenzije podatkov skoraj niso spremenile. Pri uvrščanju razredov z majhnimi razlikami z metodo L-SVM opazimo posebno nesimetrično obliko krivulje mere  $G$ .

Pri IND podatkih opazimo izrazito zmanjšanje vpliva neravnotežja ob izboru spremenljivk glede na največjo varianco (Slika 4.3.7) v primerjavi z rezultati brez zmanjšanja dimenzije podatkov (Slika 4.3.3). IND podatki se zato v primeru izbora spremenljivk glede na največjo varianco veliko bolje prilegajo realnim podatkom, kot v primeru brez zmanjšanja dimenzije podatkov. Vendar je pri uvrščanju z metodo LDA vpliv neravnotežja ob izboru spremenljivk glede na največjo varianco pri IND podatkih še vedno večji kot pri realnih podatkih. V primeru brez zmanjšanja dimenzije podatkov so bile vrednosti mere  $G$  pri IND podatkih večinoma nižje od realnih podatkov, pri izboru spremenljivk glede na največjo varianco pa so večinoma višje.

Tudi pri MVNblock podatkih lahko pri uvrščanju z metodo LDA ob uporabi izbora spremenljivk glede na največjo varianco (Slika 4.3.7) opazimo zmanjšan vpliv neravnotežja in na splošno višje vrednosti mere  $G$  kot pri rezultatih brez zmanjšanja dimenzije podatkov (Slika 4.3.3), zaradi česar se MVNblock podatki v tem primeru bolje prilegajo realnim podatkom kot v primeru brez zmanjšanja dimenzije podatkov. Pri uvrščanju z metodama CART in L-SVM se MVNblock podatki ob izboru spremenljivk glede na največjo varianco slabše prilegajo realnim podatkom v primerjavi z rezultati brez zmanjšanja dimenzije podatkov, razen v primeru uvrščanja z metodo L-SVM pri velikih razlikah med razredoma.

ABS in MVNreal podatki se ob izboru spremenljivk glede na največjo varianco (Slika 4.3.7) približno enako dobro prilegajo realnim podatkom kot v primeru brez zmanjšanja dimenzije podatkov (Slika 4.3.3). Izjema pri tem je primer uvrščanja z metodo L-SVM v primeru majhnih razlik med razredi, kjer rezultati realnih podatkov močno odstopajo od vseh ostalih (Slika 4.3.7).

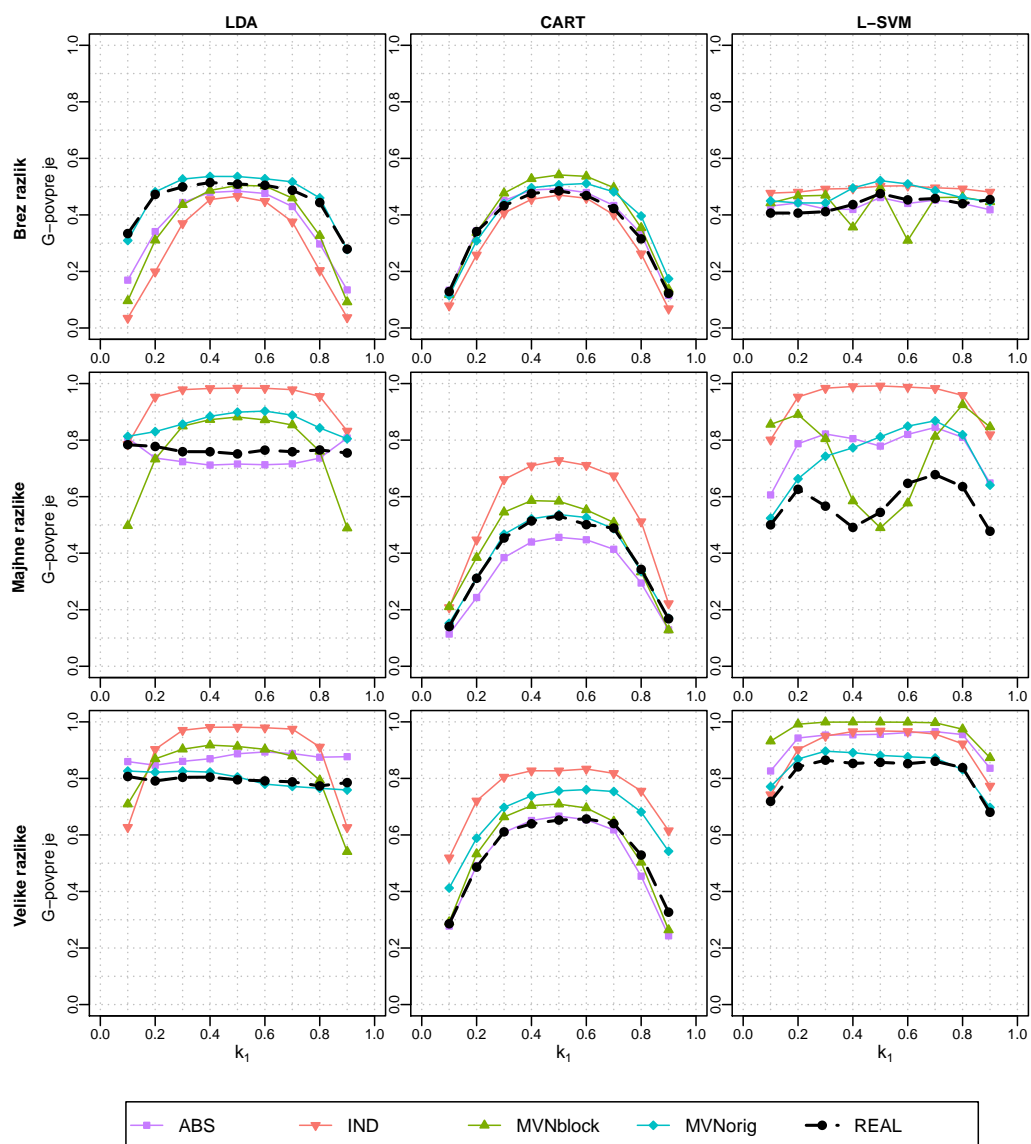
Pri uvrščanju realnih podatkov ob uporabi izbora spremenljivk glede na največjo F-statistiko (Slika 4.3.8) opazimo nižje vrednosti mere  $G$ , vpliv neravnotežja pa ostane približno enak kot v primeru brez zmanjšanja dimenzije podatkov (Slika 4.3.3), kar je podobno kot pri izboru spremenljivk glede na največjo varianco (Slika 4.3.7).

Pri IND podatkih je vpliv neravnotežja ob izboru spremenljivk glede na največjo F-statistiko (Slika 4.3.8) manjši, vrednosti mere  $G$  pa so višje kot v primeru brez zmanjšanja dimenzije podatkov (Slika 4.3.3) pri vseh obravnavanih primerih uvrščanja. Posledica tega je, da se rezultati IND podatkov v primeru izbora spremenljivk glede na največjo F-statistiko bolje prilegajo realnim podatkom kot v primeru brez zmanjšanja dimenzije podatkov. Ob tem bi z IND podatki pri izboru spremenljivk z največjo F-statistiko pri uvrščanju z LDA precenili vpliv neravnotežja glede na realne podatke, pri uvrščanju z L-SVM bi ga podcenili, pri uvrščanju s CART pa je vpliv neravnotežja približno enak kot pri realnih podatkih.

Zaradi istega razloga kot pri IND podatkih tudi pri uvrščanju MVNblock podatkov z metodo LDA ob izboru spremenljivk glede na največjo F-statistiko (Slika 4.3.8) opazimo,

da se rezultati bolje prilegajo realnim podatkom kot v primeru brez zmanjšanja dimenzije (Slika 4.3.3). Pri uvrščanju z metodo CART ob izboru spremenljivk glede na največjo F-statistiko so ostali MVNblock podatki podobni realnim. Pri uvrščanju z L-SVM pa kažejo rezultati MVNblock podatkov ob izboru spremenljivk z največjo F-statistiko manjši vpliv neravnotežja kot v primeru brez zmanjšanja dimenzije (Slika 4.3.3), kar je posledica zmanjšanja vrednosti mere G v bolj uravnoteženih primerih. Pri uvrščanju MVNblock podatkov z metodo L-SVM ob izboru spremenljivk z največjo F-statistiko so vrednosti mere G nižje od realnih in z njimi bi lahko podcenili vpliv neravnotežja realnih podatkov.

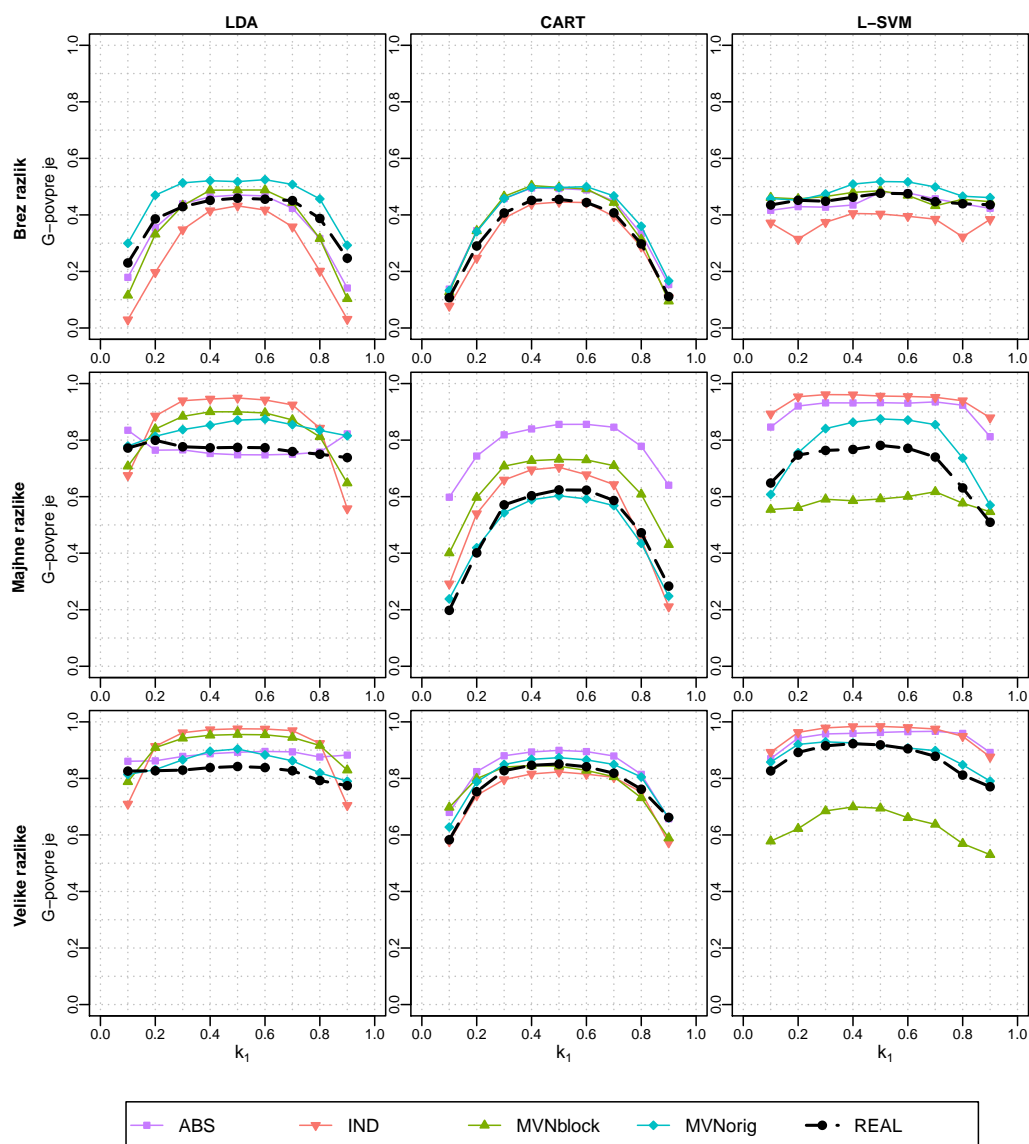
ABS in MVNorig podatki se ob izboru spremenljivk glede na največjo F-statistiko (Slika 4.3.8) približno enako dobro prilegajo realnim podatkom kot v primeru brez zmanjšanja



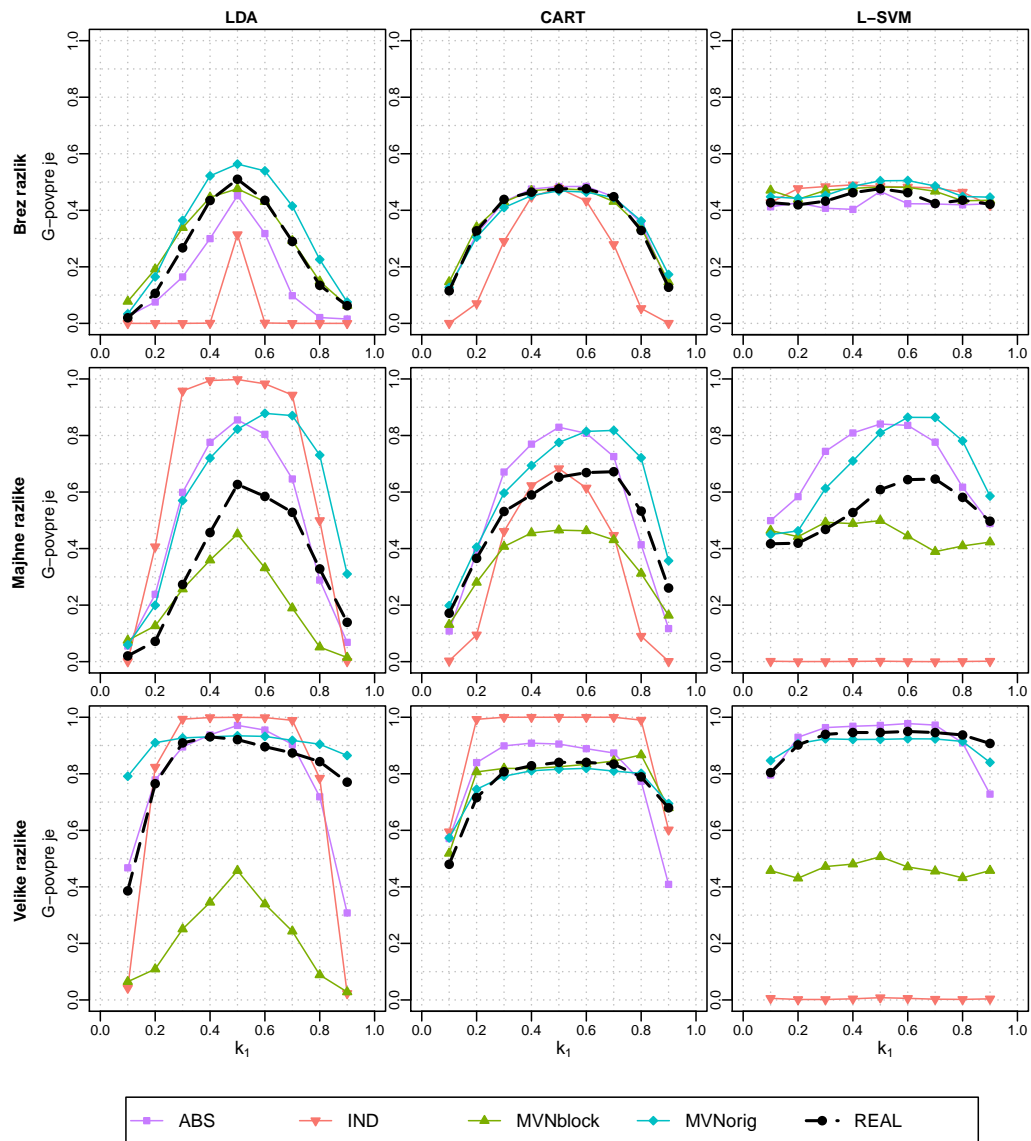
Slika 4.3.7: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM. Pri izgradnji modela uvrščanja je bilo uporabljenih 50 spremenljivk z največjo varianco.

dimenzije podatkov (Slika 4.3.3).

Pri realnih podatkih ob uporabi metode PCA za zmanjšanje dimenzije podatkov (Slika 4.3.9) rezultati uvrščanja z metodo LDA kažejo večjo občutljivost na neravnotežje kot v primeru brez zmanjšanja dimenzije podatkov (Slika 4.3.3). Pri uvrščanju z metodo CART so se ob uporabi PCA vrednosti mere G v bolj uravnoteženih primerih povečale, medtem ko se v močno neuravnoteženih primerih niso zmanjšale v primerjavi z rezultati brez zmanjšanja dimenzije podatkov. Pri uvrščanju z L-SVM ob uporabi PCA v primeru brez razlik med razredoma opazimo povečan vpliv neravnotežja, v primeru z majhnimi razlikami nižje vrednosti mere G, v primeru z velikimi razlikami se rezultati bistveno ne razlikujejo od rezultatov brez zmanjšanja dimenzije podatkov.



Slika 4.3.8: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM. Pri izgradnji modela uvrščanja je bilo uporabljenih 50 spremenljivk z največjo F-statistiko.



Slika 4.3.9: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM. Pri izgradnji modela uvrščanja so bile uporabljene glavne komponente izračunane po metodi PCA, ki so pokrile 99 % celotne variabilnosti.

Rezultati IND podatkov so tudi ob uporabi metode PCA od vseh generiranih podatkov najbolj odstopali od realnih podatkov (Slika 4.3.9). Pri uvrščanju z metodo LDA, ko so bile med razredoma razlike, opazimo visoke vrednosti mere G pri nizki stopnji neravnotežja, pri visoki stopnji neravnotežja pa so vrednosti mere G zelo nizke. Podobno opazimo tudi pri uvrščanju IND podatkov z metodo CART: vrednosti mere G so pri nizki stopnji neravnotežja višje, pri visoki stopnji neravnotežja pa nižje kot pri rezultatih brez zmanjšanja dimenzije podatkov (Slika 4.3.3). Pri uvrščanju IND podatkov z metodo L-SVM ob uporabi PCA je vpliv neravnotežja komaj opazen, vrednosti mere G so v primeru razlik med razredi zelo nizke (Slika 4.3.9).

Vpliv neravnotežja se je pri MVNblock podatkih ob uporabi PCA (Slika 4.3.9) zmanjšal v

primerjavi z uvrščanjem MVNblock podatkov brez zmanjšanja dimenzije podatkov (Slika 4.3.3), vendar so se pri tem tudi na splošno znižale vrednosti mere  $G$  pri vseh obravnavanih stopnjah neravnotežja. Zato bi ob uporabi PCA za zmanjšanje dimenzije podatkov z MVNblock podatki podcenili tako vrednosti mere  $G$  kot vpliv neravnotežja pri realnih podatkih.

Rezultati MVNorig in ABS podatkov se tudi ob uporabi PCA (Slika 4.3.9) najboljše prilagajajo realnim podatkom, čeprav je prileganje na splošno nekoliko slabše kot pri uvrščanju brez zmanjšanja dimenzije (Slika 4.3.3).

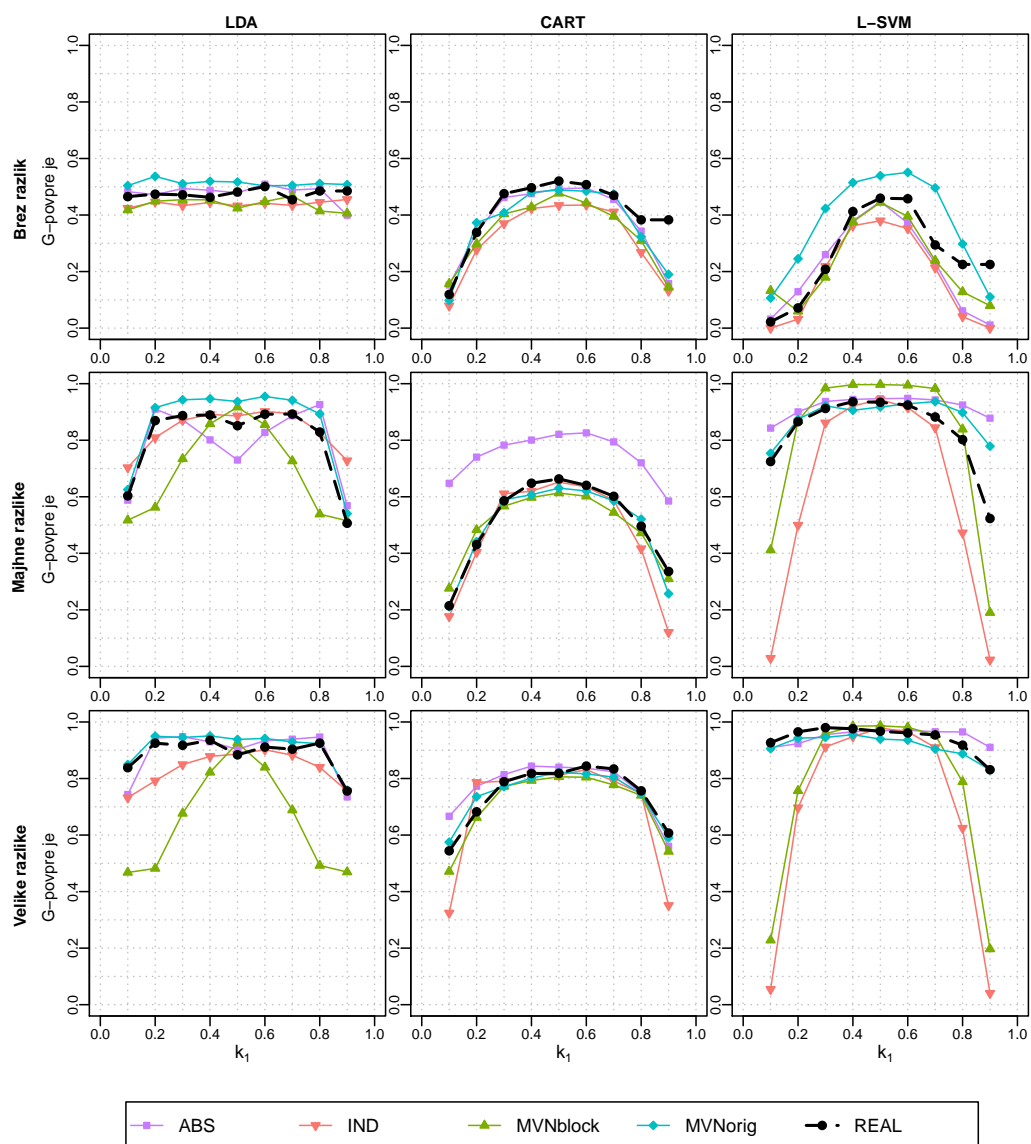
#### 4.3.2.4 Večkratno zmanjšanje večjega razreda (Slika 4.3.10)

Pri realnih podatkih ob uporabi metode MDS (Slika 4.3.10) opazimo v primerjavi z rezultati brez uporabe MDS (Slika 4.3.3) pri uvrščanju z LDA višje vrednosti mere  $G$  pri vseh stopnjah neravnotežja, razen pri najvišji, kjer so vrednosti mere  $G$  nižje kot v primeru brez MDS. Rezultati kažejo, da je MDS pri uvrščanju realnih podatkov z LDA uspešno odpravila vpliv neravnotežja, saj so nizke vrednosti mere  $G$  pri visokih stopnjah neravnotežja najverjetneje posledica majhnega števila enot. Pri uvrščanju realnih podatkov z metodo CART in L-SVM se rezultati ob uporabi MDS bistveno ne razlikujejo od tistih brez uporabe MDS, razen v primeru uvrščanja z metodo L-SVM brez razlik med razredoma, kjer ob uporabi MDS opazimo veliko večji vpliv neravnotežja.

Pri uvrščanju IND in MVNblock podatkov z metodo LDA ob uporabi MDS (Slika 4.3.10) je prileganje realnim podatkom veliko boljše kot v primeru brez uporabe MDS (Slika 4.3.3), čeprav je vpliv neravnotežja pri IND in MVNblock podatkih še vedno večji kot pri realnih podatkih. Pri uvrščanju IND in MVNblock podatkov z metodo CART ob uporabi MDS je prileganje realnim podatkom približno enako kot brez uporabe MDS. Rezultati uvrščanja IND in MVNblock podatkov z metodo L-SVM pa se ob uporabi MDS slabše prilagajajo realnim podatkom kot brez uporabe MDS, saj IND in MVNblock podatki ob uporabi MDS kažejo veliko večjo občutljivost na neravnotežje.

Podobno kot v primeru brez uporabe MDS (Slika 4.3.3) se tudi ob uporabi MDS (Slika 4.3.10) rezultati uvrščanja MVNorig najboljše prilagajajo realnim podatkom, drugo najboljše prileganje realnim podatkom pa lahko opazimo pri ABS podatkih.

Pri uvrščanju z metodo L-SVM ob uporabi metode MDS nismo mogli posebej prilagoditi praga za uvrščanje, saj so pri metodi MDS množice, ki vstopajo v model uvrščanja uravnotežene. Če primerjamo rezultate uvrščanja z metodo L-SVM brez uporabe MDS (Slika 4.3.3) z rezultati z uporabo MDS (Slika 4.3.10) opazimo, da je prilagoditev praga učinkoviteje zmanjšala vpliv neravnotežja, kar je še posebej očitno v primeru brez razlik med razredoma za vse vrste podatkov, v primerih z razlikami med razredoma pa le pri uvrščanju IND podatkov.



Slika 4.3.10: Povprečne vrednosti mere G pri uvrščanju v dva razreda z metodami LDA, CART in L-SVM. Pri izgradnji modela uvrščanja je bilo uporabljeno večkratno zmanjšanje večjega razreda v 100 ponovitvah. Simulacije so bile v tem primeru izvedene le v 100 ponovitvah.

### 4.3.2.5 Povzetek rezultatov pri simulacijah uvrščanja v dva razreda

V tem razdelku smo skušali povzeti rezultate simulacij uvrščanja v dva razreda. Za vsako metodo uvrščanja in načrtovalno strategijo (uporabljeno predobdelavo, metodo za zmanjšanje dimenzije podatkov ali MDS) so predstavljeni rezultati uvrščanja realnih podatkov (Tabela 4.3.1) in podobnost med rezultati uvrščanja umetno generiranih podatkov z rezultati uvrščanja realnih podatkov (Tabela 4.3.2).

Metoda uvrščanja	Strategija	Brez razlik		Majhne razlike		Velike razlike	
		Mera G	Razlika	Mera G	Razlika	Mera G	Razlika
		pri $k_1 = 0,5$	med $k_1 = 0,5$ in $k_1 = 0,2$	pri $k_1 = 0,5$	med $k_1 = 0,5$ in $k_1 = 0,2$	pri $k_1 = 0,5$	med $k_1 = 0,5$ in $k_1 = 0,2$
LDA	brez	0,47	0,09	0,84	0,01	0,88	0,03
LDA	izbor – varianca	0,51	0,04	0,75	0,03	0,79	0,00
LDA	izbor – F-statistika	0,46	0,07	0,77	0,03	0,84	0,02
LDA	PCA	0,51	0,40	0,63	0,55	0,92	0,16
LDA	predobdelava – SNV	0,52	0,03	0,86	0,05	0,97	0,04
LDA	predobdelava – 1. odvod	0,48	0,06	0,84	0,01	0,88	0,02
LDA	MDS	0,48	0,01	0,85	0,02	0,88	0,04
CART	brez	0,46	0,15	0,65	0,20	0,82	0,12
CART	izbor – varianca	0,49	0,14	0,53	0,22	0,65	0,17
CART	izbor – F-statistika	0,45	0,16	0,62	0,22	0,85	0,10
CART	PCA	0,48	0,15	0,65	0,29	0,84	0,12
CART	predobdelava – SNV	0,48	0,15	0,87	0,12	0,92	0,09
CART	predobdelava – 1. odvod	0,47	0,14	0,95	0,04	0,95	0,02
CART	MDS	0,52	0,18	0,66	0,23	0,82	0,14
L-SVM	brez	0,49	0,02	0,92	0,01	0,96	0,02
L-SVM	izbor – varianca	0,48	0,07	0,54	0,08	0,86	0,02
L-SVM	izbor – F-statistika	0,48	0,02	0,78	0,03	0,92	0,03
L-SVM	PCA	0,48	0,06	0,61	0,19	0,95	0,04
L-SVM	predobdelava – SNV	0,48	0,01	0,89	0,03	0,99	0,00
L-SVM	predobdelava – 1. odvod	0,48	0,01	0,95	0,00	0,98	0,00
L-SVM	MDS	0,46	0,39	0,93	0,07	0,97	0,00

Tabela 4.3.1: **Povzetek simulacij realnih podatkov pri uvrščanju v dva razreda.** V stolpcih **Mera G** so prikazane vrednosti mere G v uravnoveženem primeru ( $k_1 = 0,5$ ). V stolpcih **Razlika** pa so prikazane absolutne razlike med vrednostma mere G pri  $k_1 = 0,5$  in  $k_1 = 0,2$  – večja razlika predstavlja večji vpliv neravnotežja. Najtemnejša barva v stolpcih **Mera G** predstavlja najvišjo vrednost mere G, v stolpcih **Razlika** pa najmanjšo razliko.

Pri metodi LDA so se na splošno kot najboljše strategije pri uvrščanju izkazale uporaba MDS, predobdelava s 1. odvodom in predobdelava z metodo SNV. Pri metodi CART in L-SVM so bili najboljši rezultati tudi pri predobdelavi s 1. odvodom, nekoliko slabši pa pri predobdelavi s SNV (Tabela 4.3.2).



metoda uvrščanja	strategija	Brez razlik				Majhne razlike				Velike razlike			
		IND	MVN block	ABS	MVN orig	IND	MVN block	ABS	MVN orig	IND	MVN block	ABS	MVN orig
LDA	brez	0,29	0,09	0,04	0,05	0,44	0,21	0,08	0,07	0,45	0,20	0,03	0,03
	izbor – varianca	0,16	0,09	0,09	0,02	0,17	0,13	0,04	0,09	0,17	0,11	0,08	0,02
	izbor – F-statistika	0,12	0,05	0,04	0,07	0,14	0,09	0,03	0,07	0,12	0,10	0,06	0,03
	PCA	0,22	0,03	0,09	0,07	0,34	0,15	0,17	0,24	0,18	0,60	0,09	0,09
	predobdelava – SNV	0,37	0,08	0,08	0,03	0,55	0,11	0,09	0,09	0,67	0,17	0,01	0,02
	predobdelava – 1. odvod	0,23	N.P.	0,04	0,05	0,52	N.P.	0,07	0,07	0,58	N.P.	0,04	0,03
	MDS	0,04	0,04	0,02	0,04	0,05	0,13	0,06	0,05	0,05	0,25	0,03	0,02
		0,02	0,08	0,04	0,04	0,02	0,04	0,24	0,03	0,07	0,04	0,03	0,01
CART	brez	0,04	0,04	0,01	0,04	0,16	0,05	0,06	0,01	0,21	0,17	0,02	0,12
	izbor – varianca	0,02	0,04	0,04	0,05	0,08	0,14	0,29	0,03	0,03	0,21	0,05	0,03
	izbor – F-statistika	0,13	0,01	0,01	0,02	0,17	0,15	0,12	0,10	0,17	0,06	0,09	0,03
	PCA	0,04	0,02	0,03	0,02	0,32	0,22	0,07	0,04	0,24	0,19	0,02	0,04
	predobdelava – SNV	0,02	0,02	0,01	0,06	0,53	0,06	0,02	0,06	0,17	0,44	0,03	0,05
	predobdelava – 1. odvod	0,10	0,08	0,04	0,05	0,05	0,04	0,23	0,03	0,07	0,26	0,04	0,02
	MDS	0,01	0,02	0,02	0,01	0,05	0,22	0,08	0,04	0,05	0,04	0,01	0,01
		0,05	0,05	0,02	0,04	0,37	0,21	0,20	0,17	0,09	0,17	0,11	0,03
L-SVM	brez	0,08	0,01	0,01	0,03	0,23	0,13	0,20	0,08	0,08	0,18	0,07	0,02
	izbor – varianca	0,03	0,02	0,02	0,03	0,53	0,11	0,16	0,15	0,92	0,29	0,04	0,03
	izbor – F-statistika	0,07	0,03	0,02	0,01	0,05	0,21	0,13	0,06	0,12	N.P.	0,01	0,03
	PCA	0,02	0,01	0,01	0,01	0,30	0,05	0,03	0,03	0,45	N.P.	0,02	0,04
	predobdelava – SNV	0,09	0,06	0,08	0,13	0,22	0,12	0,08	0,06	0,26	N.P.	0,03	0,03
	predobdelava – 1. odvod												
	MDS												

Tabela 4.3.2: **Podobnost rezultatov uvrščanja med generiranimi in realnimi podatki.** Kot mera podobnosti je izračunana povprečna absolutna razlika med G-vrednostmi posameznih generiranih in realnih podatkov pri vseh obravnavanih stopnjah neravnotežja, ki so prikazane na slikah v razdelku 4.3.2. Z barvami so vrednosti v tabeli rangirane: temnejša barva predstavlja večjo podobnost z realnimi podatki.

V primeru dveh enakih razredov so se realnim podatkom na splošno (ne glede na metodo uvrščanja in strategijo) najbolj približali ABS podatki. Če gledamo ločeno po metodah uvrščanja, vidimo da so realnim podatkom pri uvrščanju z LDA najbližji MVNorig podatki, pri uvrščanju s CART ABS podatki, pri uvrščanju z L-SVM pa spet ABS podatki (Tabela 4.3.2).

V primeru uvrščanja v dva razreda z majhnimi in velikimi razlikami so se realnim podatkom najbolj približali MVNorig podatki, sledili pa so jim ABS podatki. Z realnimi podatki so bili najmanj usklajeni IND podatki (Tabela 4.3.2).

### 4.3.3 Uvrščanje v tri razrede

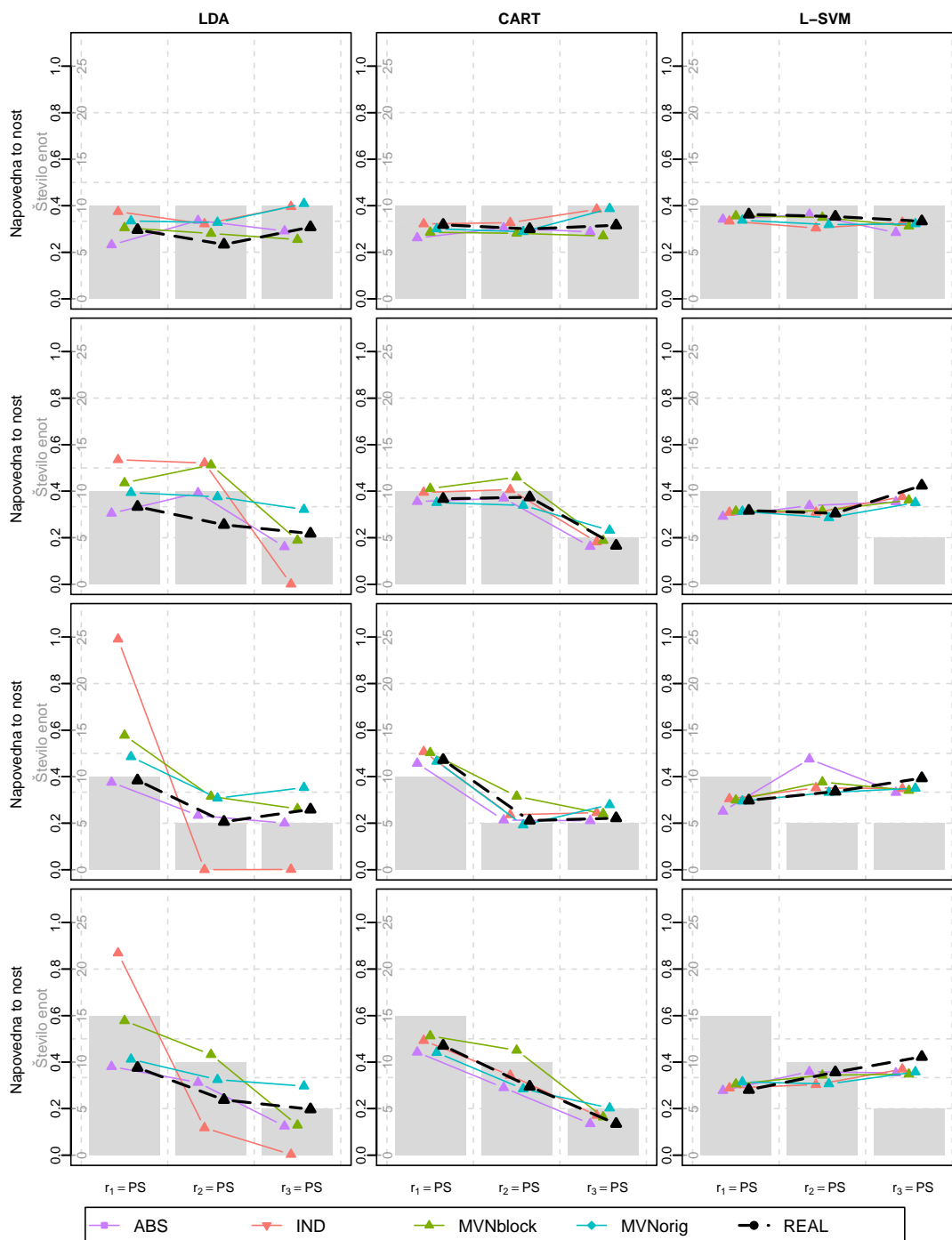
S simulacijami uvrščanja v tri razrede smo želeli preučiti vpliv predobdelave (SNV in 1. odvod), zmanjšanja dimenzije podatkov (izbor 50 spremenljivk z največjo varianco, izbor 50 spremenljivk z največjo F-statistiko in metoda PCA) in metode MDS na uvrščanje NIRS podatkov. Ob tem smo želeli preučiti, kako dobro se rezultati umetno generiranih podatkov približajo realnim.

#### 4.3.3.1 Brez predobdelave in brez zmanjšanja dimenzije podatkov (Slike 4.3.11, 4.3.12, 4.3.13, 4.3.14 in 4.3.15)

Pri uvrščanju v tri razrede, med katerimi ni razlik, pričakujemo, da bodo enote v razrede uvrščene popolnoma naključno, torej da bo napovedna točnost posameznega razreda enaka  $1/3$ . To lahko opazimo v uravnoteženem primeru pri uvrščanju realnih podatkov z metodo L-SVM brez uporabe predobdelave ali zmanjšanja dimenzije podatkov (Slika 4.3.11). Pri uvrščanju realnih podatkov z metodama LDA in CART v uravnoteženem primeru pa so bile napovedne točnosti vseh treh razredov nekoliko nižje od  $1/3$ . V primeru dveh velikih in enega majhnega razreda opazimo vpliv neravnotežja pri uvrščanju realnih podatkov z metodo LDA in CART, kjer je bila napovedna točnost majhnega razreda nižja od napovedne točnosti v uravnoteženem primeru, napovedna točnost velikih razredov pa višja od napovedne točnosti v uravnoteženem primeru. Razlika med napovednimi točnostmi posameznih razredov v uravnoteženem in neuravnoteženem primeru je bila pri uvrščanju s CART večja kot pri uvrščanju z LDA, kar kaže na to, da je bil vpliv neravnotežja pri CART večji kot pri LDA. V primeru dveh velikih in enega majhnega razreda pri uvrščanju realnih podatkov z metodo L-SVM pa nasprotno kot pri uvrščanju z LDA in CART opazimo, da je bila napovedna točnost majhnega razreda višja od napovednih točnosti velikih razredov. Vzrok za to je najverjetneje uporaba prilagojenega praga za uvrščanje (metoda je opisana v razdelku 3.3.2.3), ki je bil v tem primeru nekoliko "preveč" prilagojen. V primeru, ko smo imeli en velik in dva majhna razreda, so opažanja pri uvrščanju realnih podatkov podobna, kot v primeru dveh velikih in enega majhnega razreda: pri LDA in CART je imel večji razred višjo napovedno točnost kot manjša dva, pri L-SVM pa sta imela manjša dva nekoliko višjo napovedno točnost kot večji razred. Pri tem lahko pri uvrščanju z LDA in CART opazimo še, da je imel večji razred napovedno točnost višjo, kot je bila napovedna točnost velikih razredov v primeru dveh velikih in enega majhnega razreda, prav tako je bila napovedna točnost manjših dveh razredov (v primeru dveh malih in enega velikega razreda) nekoliko višja kot napovedna točnost majhnega razreda v primeru dveh velikih in enega majhnega razreda. Tudi pri uvrščanju realnih podatkov v primeru treh različno velikih razredov, med katerimi ni razlik (Slika 4.3.11), opazimo podoben vpliv neravnotežja kot v prejšnjih primerih: pri LDA in CART so se napovedne točnosti z večanjem velikosti razreda zviševale, pri L-SVM pa so se zniževale.

Ko je bila uporabljena metoda LDA za uvrščanje umetno generiranih podatkov v tri razrede, med katerimi ni bilo razlik (Slika 4.3.11), opazimo, da so se rezultati uvrščanja IND podatkov najbolj razlikovali od rezultatov realnih podatkov, saj so IND podatki pokazali veliko večji vpliv neravnotežja kot realni podatki. Pri uvrščanju z metodama CART in L-SVM pa so se rezultati IND podatkov dobro prilegali realnim podatkom.

Podobno kot pri IND podatkih lahko tudi pri MVNblock podatkih ob uporabi metode LDA za uvrščanje v tri razrede, med katerimi ni bilo razlik (Slika 4.3.11), opazimo večji



Slika 4.3.11: **PS-PS-PS**. Povprečne napovedne točnosti po razredih, izračunane pri uvrščanju v tri razrede, med katerimi ni bilo razlik (enote v vseh treh razredih so bile iz skupine PS).

vpliv neravnotežja kot pri realnih podatkih, čeprav je bil vpliv neravnotežja vseeno manjši kot pri IND podatkih. Pri uvrščanju z metodo CART so se rezultati MVNblock podatkov bolje prilegali rezultatom realnih podatkov kot pri uvrščanju z LDA, le napovedne točnosti sredinskega razreda so nekoliko precenile realne napovedne točnosti. Pri uvrščanju z metodo

L-SVM so se rezultati MVNblock podatkov dobro prilegali realnim.

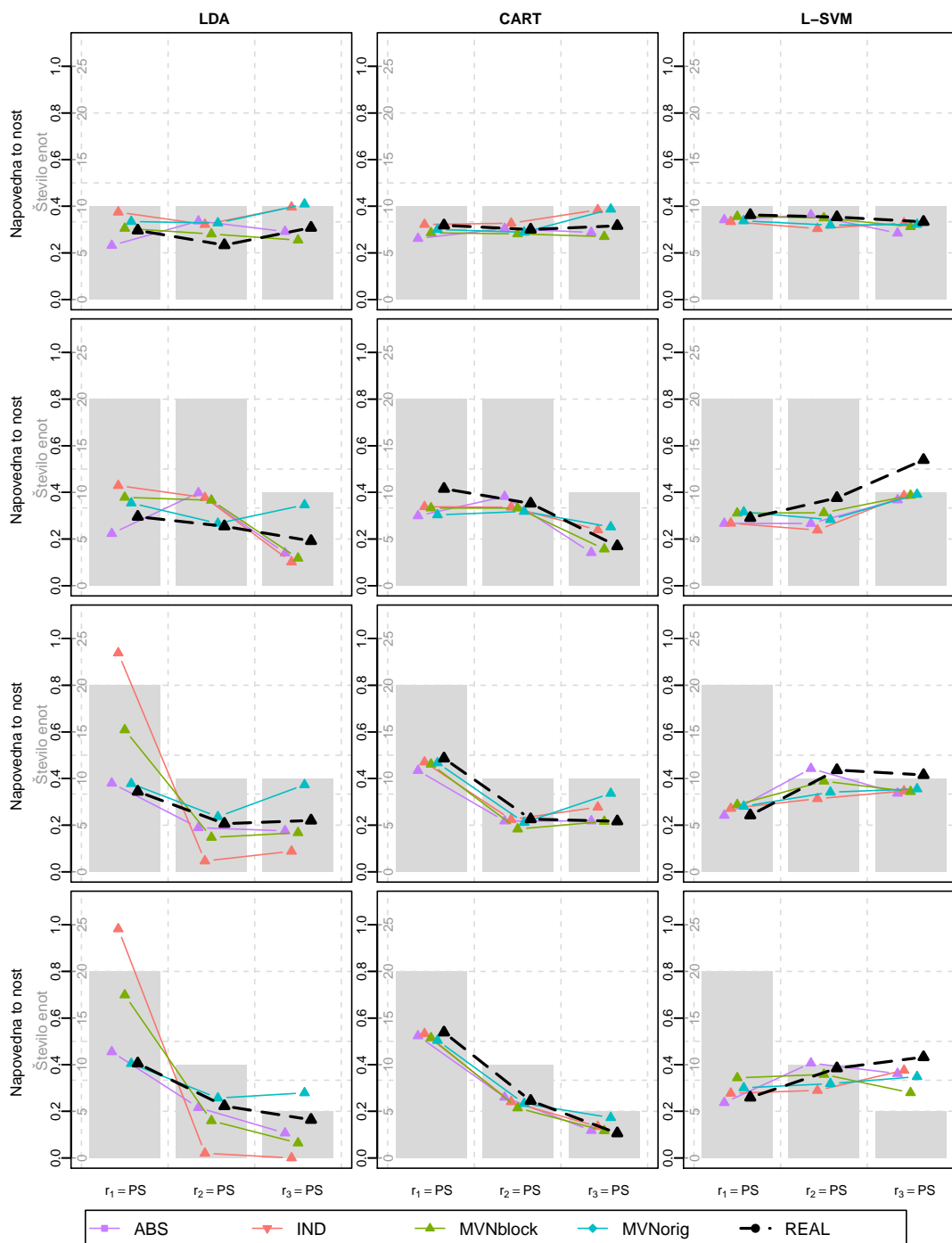
Ob uporabi metode LDA za uvrščanje ABS podatkov v tri razrede, med katerimi ni bilo razlik (Slika 4.3.11), opazimo, da so bile napovedne točnosti velikih razredov primerljive z napovednimi točnostmi velikih razredov pri realnih podatkih, medtem ko so bile napovedne točnosti najmanjšega razreda nekoliko nižje kot pri realnih podatkih. Pri uvrščanju z metodo CART so se rezultati ABS podatkov zelo dobro ujemali z rezultati realnih podatkov. Podobno lahko opazimo tudi pri uvrščanju z L-SVM, kjer je izjema le primer uvrščanja z enim velikim in dvema majhnima razredoma, kjer je imel sredinski razred pri ABS podatkih veliko višjo, ostala dva razreda pa nekoliko nižjo napovedno točnost od realnih rezultatov.

Ko je bila uporabljena metoda LDA za uvrščanje MVNorig podatkov v tri razrede, med katerimi ni bilo razlik (Slika 4.3.11), opazimo, da bi z njimi v vseh primerih razen v uravnoteženem primeru precenili realne napovedne točnosti posameznih razredov. Podobno bi z MVNorig podatki nekoliko precenili realne napovedne točnosti razredov tudi pri uvrščanju s CART, medtem ko se pri uvrščanju z L-SVM rezultati MVNorig podatkov dobro prilegajo realnim rezultatom.

Ko smo povečali število enot po razredih (Slika 4.3.12), pri čemer med razredi še vedno ni bilo razlik, pri uvrščanju realnih podatkov z metodama LDA in CART nismo opazili bistvenih razlik ne glede na stopnjo neravnotežja v primerjavi z uvrščanjem, kjer so bili razredi manjši (Slika 4.3.11). Samo v primeru treh različno velikih razredov smo pri večjih razredih (Slika 4.3.12) opazili večji vpliv neravnotežja kot pri manjših razredih (Slika 4.3.11), kar pa ni bila posledica večjih razredov, pač pa večjega neravnotežja med razredi, saj se v primeru treh različno velikih razredov na Slikah 4.3.12 in 4.3.11 razlikujeta le velikosti največjega razreda (20 in 10). Pri uvrščanju realnih podatkov z metodo L-SVM in večjem številu enot po razredih (Slika 4.3.12) opazimo, da so bile napovedne točnosti večjih razredov še nižje, napovedne točnosti manjših razredov pa višje kot v primeru z manj enotami po razredih (Slika 4.3.11).

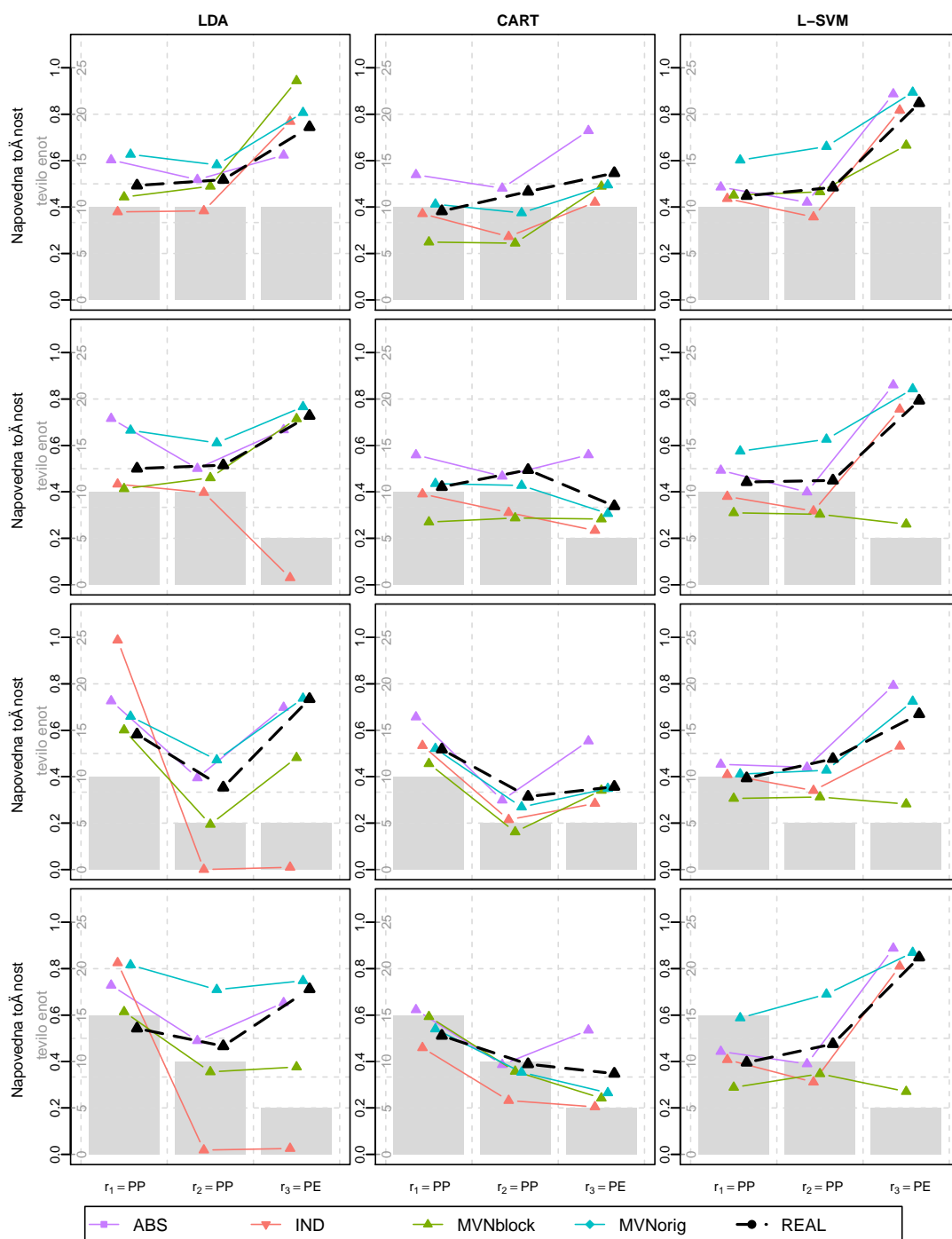
Rezultati generiranih podatkov se v primeru večjih razredov (Slika 4.3.12) podobo prilegajo rezultatom realnih podatkov kot v primeru, ko je bilo število enot po razredih manjše (Slika 4.3.11).

Napovedne točnosti posameznih razredov realnih podatkov so bile pri uvrščanju z metodo LDA v primeru, ko se je en razred z manjšimi razlikami razlikoval od drugih dveh (PP-PP-PE, Slika 4.3.13) na splošno višje kot v primeru brez razlik med razredi (PS-PS-PS, Slika 4.3.11). Razred PE, ki je bil različno izražen od drugih dveh, je imel v vseh primerih neravnotežja najvišje napovedne točnosti. Tudi pri uvrščanju realnih podatkov s CART v primeru PP-PP-PE opazimo višje napovedne točnosti v primerjavi s primerom, ko so bili vsi razredi enako izraženi (PS-PS-PS, Slika 4.3.11), vendar je bila napovedna točnost razreda PE pri uvrščanju s CART le v uravnoteženem razredu višja od napovednih točnosti ostalih dveh razredov. V ostalih obravnavanih primerih neravnotežja so bile najvišje napovedne točnosti največjega razreda, kar kaže na to, da razlike med razredi niso odpravile vpliva neravnotežja tako učinkovito kot pri uvrščanju z LDA. Pri uvrščanju realnih podatkov z L-SVM v primeru PP-PP-PE (Slika 4.3.13) so opažanja podobna kot pri metodi LDA: napovedne točnosti vseh razredov so višje v primerjavi z napovednimi točnostmi, kjer med razredi ni bilo razlik, najvišje napovedne točnosti pa ima razred, ki je od drugih dveh različno izražen – razred PE.



Slika 4.3.12: **PS-PS-PS (večji razredi)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, med katerimi ni bilo razlik (enote v vseh treh razredih so bile iz skupine PS). Število enot v posameznem razredu je bilo večje kot na prejšnji sliki (Slika 4.3.11).

Pri uvrščanju IND in MVNblock podatkov z metodo LDA v primeru, ko se je en razred z manjšimi razlikami razlikoval od drugih dveh (PP-PP-PE, Slika 4.3.13), razlika med razredi ne odpravi povsem vpliva neravnotežja, kar povzroči, da so napovedne točnosti



Slika 4.3.13: **PP-PP-PE**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer med prvima dvema razredoma ni bilo razlik (skupina PP), tretji razred pa se je z majhnimi razlikami razlikoval od prvih dveh (skupina PE).

različno izraženega razreda pri IND in MVNblock podatkih veliko nižje od realnih vrednosti. Pri uvrščanju MVNorig in ABS podatkov z metodo LDA opazimo dobro prileganje realnim rezultatom, čeprav bi z metodo MVNorig v vseh primerih nekoliko precenili realne rezultate.

Pri uvrščanju umetno generiranih podatkov z metodo CART v primeru, ko se en razred z manjšimi razlikami razlikuje od drugih dveh (Slika 4.3.13), so se realnim podatkom najboljše prilegali MVNorig podatki. ABS podatki so v vseh primerih precenili napovedne točnosti različno izraženega razreda, medtem ko so jih MVNblock in IND podatki nekoliko podcenili.

Pri uvrščanju umetno generiranih podatkov z metodo L-SVM v primeru, ko se en razred z manjšimi razlikami razlikuje od drugih dveh (Slika 4.3.13), so se rezultati pri pri vseh umetno generiranih podatkih sorazmerno dobro ujemali z rezultati realnih podatkov, izjema so bili le MVNblock podatki, pri katerih bi napovedne točnosti različno izraženega razreda močno podcenile realne vrednosti.

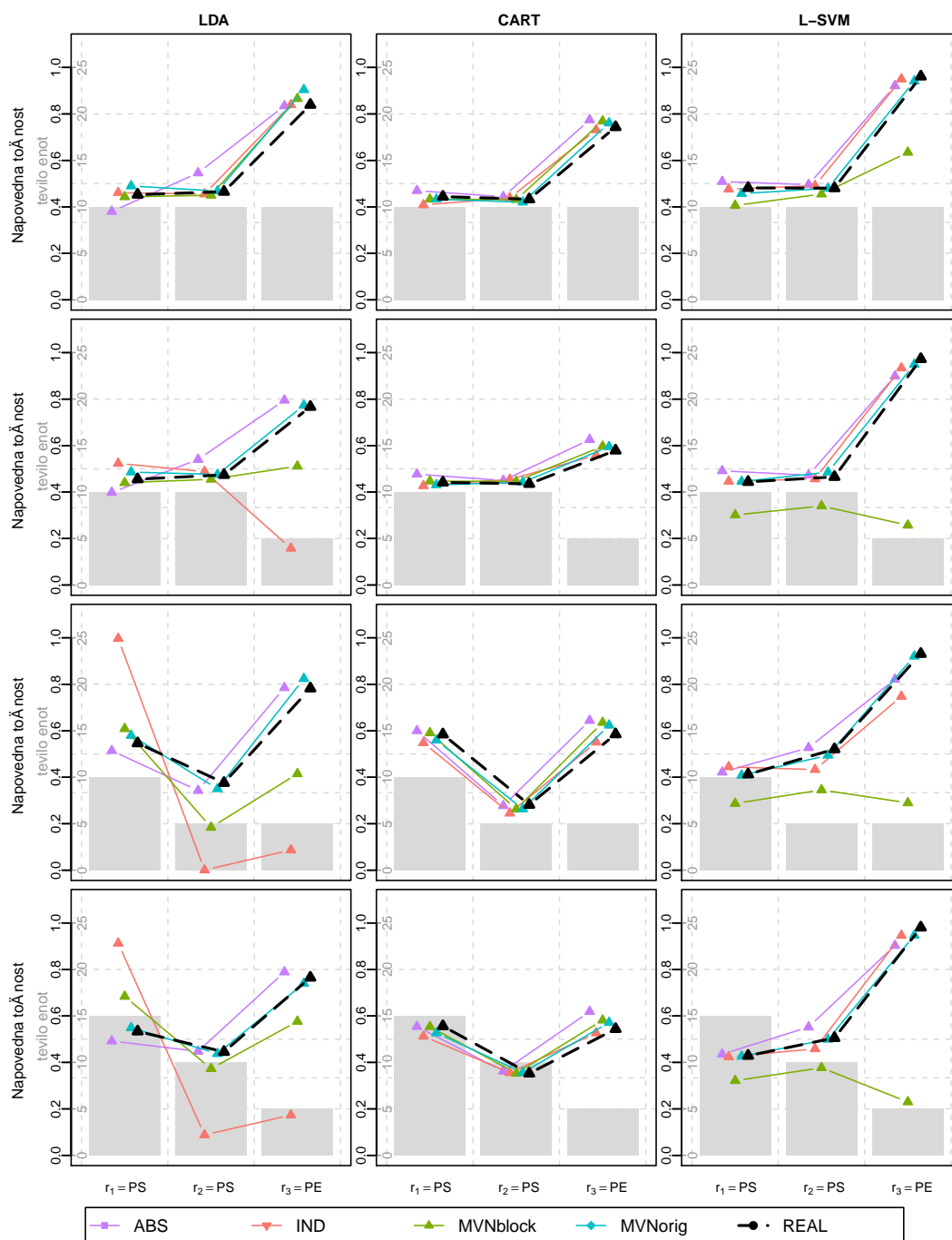
Pri uvrščanju realnih podatkov v tri razrede, kjer se je en razred z velikimi razlikami razlikoval od drugih dveh (PS-PS-PE, Slika 4.3.14), so opažanja pri uvrščanju z LDA in L-SVM podobna kot v primeru uvrščanja, kjer se je en razred z majhnimi razlikami razlikoval od drugih dveh (Slika 4.3.13): napovedne točnosti vseh razredov so višje kot v primeru brez razlik med razredi (Slika 4.3.11), med vsemi razredi pa ima najvišjo napovedno točnost razred, ki je različno izražen pa tudi, če je najmanjši. Pri uvrščanju z metodo CART v primeru, ko se en razred z velikimi razlikami razlikuje od drugih dveh (Slika 4.3.14), opazimo, da so se napovedne točnosti različno izraženega razreda močno zvišale v primerjavi s primerom, kjer se je en razred z majhnimi razlikami razlikoval od drugih dveh (PP-PP-PE, Slika 4.3.13). Ob tem lahko pri obravnavanih neuravnoteženih primerih opazimo, da se je najmanjši razred dovolj razlikoval od ostalih dveh, da vpliv neravnotežja pri tem razredu ni bil več izrazit.

Pri uvrščanju umetno generiranih podatkov v primeru, ko se en razred z velikimi razlikami razlikuje od ostalih dveh (Slika 4.3.14) ob uporabi metode LDA opazimo, da so v neuravnoteženih primerih podatki MVNblock in IND močno podcenili realne vrednosti manjših razredov. V vseh ostalih primerih uvrščanja so bili rezultati vseh umetno generiranih podatkov primerljivi z realnimi rezultati, izjema pri tem so bili le MVNblock podatki pri uvrščanju z L-SVM, ki kažejo v splošnem slabše napovedne točnosti in večji vpliv neravnotežja kot realni rezultati.

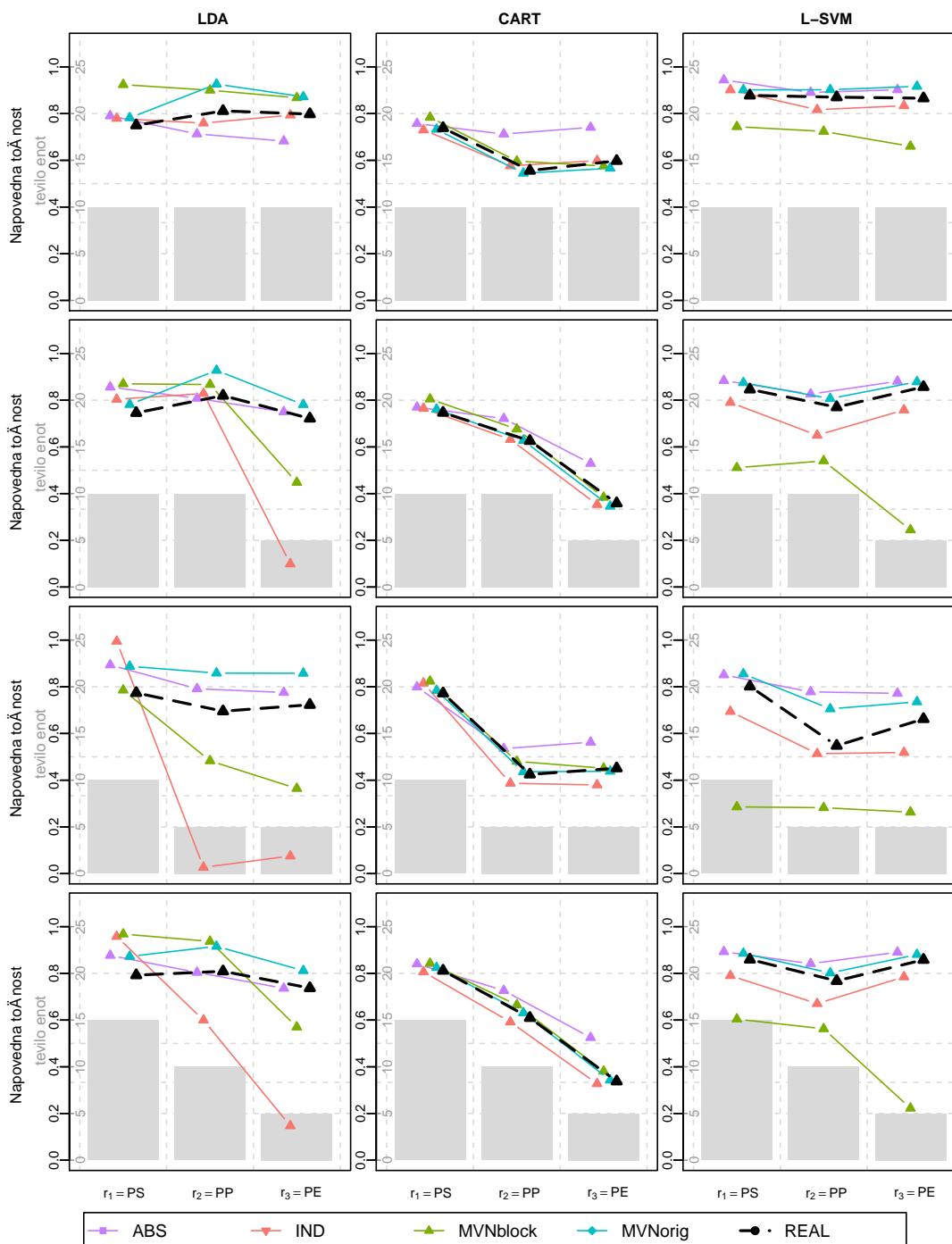
Pri uvrščanju realnih podatkov v tri razrede v primeru, ko se vsi trije razredi med seboj razlikujejo (Slika 4.3.15), so bile pri vseh treh metodah uvrščanja (LDA, CART in L-SVM) na splošno višje napovedne točnosti posameznih razredov pri vseh opazovanih primerih neravnotežja v primerjavi z rezultati, kjer med razredi ni bilo razlik (Slika 4.3.11) ali se je le eden od razredov razlikoval od drugih dveh (Slika 4.3.13 in Slika 4.3.14). Ob tem pri metodah LDA in L-SVM skoraj ne opazimo vpliva neravnotežja, medtem ko je le-ta pri uvrščanju z metodo CART zelo izrazit.

Pri uvrščanju umetno generiranih podatkov v primeru, ko se vsi razredi med seboj razlikujejo (Slika 4.3.15), so se pri uvrščanju z metodo LDA od realnih podatkov izrazito razlikovali rezultati MVNblock in IND podatkov, pri uvrščanju z L-SVM pa rezultati MVNblock podatkov; v vseh ostalih obravnavanih primerih uvrščanja so se rezultati vseh umetno generiranih podatkov sorazmerno dobro prilegali realnim vrednostim. Pri metodah CART in L-SVM so se rezultati uvrščanja MVNorig podatkov najbolj približali rezultatom realnih podatkov, pri uvrščanju z LDA pa so se jim najbolj približali rezultati ABS podatkov.





Slika 4.3.14: **PS-PS-PE**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer med prvima dvema razredoma ni bilo razlik (skupina PS), tretji razred pa se je z velikimi razlikami razlikoval od prvih dveh (skupina PE).



Slika 4.3.15: **PS-PP-PE**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer so bile enote v prvem razredu iz skupine PS, v drugem razredu iz skupine PP in v tretjem razredu iz skupine PE. Razlike med skupinama PP in PE so bile manjše kot med PP in PS ali PE in PS.

#### 4.3.3.2 Predobdelave spektrov, zmanjšanje dimenzije podatkov in večkratno zmanjšanje večjega razreda

Rezultati uvrščanja v tri razrede ob uporabi predobdelave, zmanjšanja dimenzije podatkov in metode MDS so prikazani v dodatku D, v tem razdelku pa je le njihov povzetek.

Ko smo pri predobdelavi realnih podatkov uporabili metodo SNV, smo pri uvrščanju z vsemi tremi metodami uvrščanja (LDA, CART in L-SVM) opazili višje napovedne točnosti posameznih razredov, pri metodi CART pa smo ob tem opazili tudi manjši vpliv neravnotežja kot v primeru, ko predobdelava ni bila uporabljena.

Ob uporabi SNV predobdelave pri uvrščanju umetno generiranih podatkov so se rezultati ABS in MVNorig podatkov pri vseh metodah uvrščanja podobno prilegali realnim podatkom kot brez uporabe predobdelave. Pri rezultatih uvrščanja IND podatkov smo pri metodi LDA opazili še večji vpliv neravnotežja kot v primeru brez uporabe predobdelave, pri metodah CART in L-SVM pa med rezultati uvrščanja IND podatkov brez predobdelave in s SNV predobdelavo nismo opazili razlik. Vse to je povzročilo, da so se rezultati pri IND podatkih ob uporabi SNV še bolj razlikovali od realnih podatkov kot v primeru brez uporabe predobdelave. Pri MVNblock podatkih opazimo, da se rezultati uvrščanja z metodo LDA ob uporabi SNV nekoliko bolje prilegajo realnim vrednostim kot brez uporabe predobdelave, čeprav je bil vpliv neravnotežja pri MVNblock podatkih še vedno večji kot pri realnih podatkih. Pri CART se vrednosti MVNblock podatkov ob uporabi SNV niso bistveno spremenile v primerjavi z rezultati brez uporabe predobdelave, pri L-SVM se rezultati MVNblock podatkov realnim podatkom ob uporabi SNV še slabše prilegajo kot brez uporabe predobdelave.

Ob uporabi 1. odvoda pri predobdelavi realnih podatkov smo opazili, da se rezultati uvrščanja z metodo LDA niso razlikovali od primera, kjer predobdelave nismo uporabili. Uporaba 1. odvoda je izboljšala napovedne točnosti posameznih razredov in zmanjšala vpliv neravnotežja pri uvrščanju realnih podatkov z metodama CART in L-SVM v primerjavi z rezultati, kjer predobdelava ni bila uporabljena.

Ob uporabi 1. odvoda pri uvrščanju umetno generiranih podatkov so bili rezultati MVNorig in ABS podatkov blizu realnim, medtem ko so rezultati pri MVNblock in IND močno odstopali od realnih vrednosti.

Ko smo modele uvrščanja zgradili na 50 spremenljivkah z najvišjo varianco, so bile pri realnih podatkih napovedne točnosti posameznih razredov pri vseh obravnavanih metodah uvrščanja nižje kot v primeru brez zmanjšanja dimenzije podatkov.

Pri umetno generiranih podatkih smo ob izboru 50 spremenljivk z največjo varianco pri uvrščanju z LDA opazili, da so rezultati pri IND in MVNblock bližje realnim podatkom kot v primeru brez zmanjšanja dimenzije podatkov, izbor spremenljivk je namreč pri IND in MVNblock podatkih povišal napovedne točnosti posameznih razredov in zmanjšal vpliv neravnotežja v primerjavi z rezultati brez zmanjšanja dimenzije podatkov. Pri uvrščanju umetno generiranih podatkov z metodo CART so bili rezultati ob izboru spremenljivk z največjo varianco podobni kot brez zmanjšanja dimenzije podatkov, samo pri IND podatkih smo opazili večje odstopanje od realnih vrednosti. Pri uvrščanju z L-SVM so ob izboru spremenljivk z največjo varianco vsi umetno generirani podatki, razen MVNblock podatkov, precenili realne vrednosti, pri tem so od realnih podatkov najbolj odstopali rezultati IND podatkov. Rezultati vseh umetno generiranih podatkov so bili z realnimi vrednostmi manj

skladni kot v primeru brez zmanjšanja dimenzije podatkov.

Ko smo pri uvrščanju realnih podatkov za zmanjšanje dimenzije uporabili izbor spremenljivk z največjo F-statistiko, smo pri metodah LDA in L-SVM opazili nekoliko nižje napovedne točnosti v primerjavi z rezultati brez zmanjšanja dimenzije, na CART pa izbor spremenljivk z največjo F-statistiko ni bistveno vplival.

Ob izboru spremenljivk z največjo F-statistiko pri uvrščanju umetno generiranih podatkov smo pri LDA opazili, da so se realnim rezultatom najbolj prilegali rezultati ABS podatkov, sledili so jim rezultati MVNorig podatkov, ki pa so pri vseh obravnavanih primerih neravnotežja precenili realne vrednosti. Pri uvrščanju s CART smo ob izboru spremenljivk glede na največjo F-statistiko opazili, da so se realnim vrednostim najbolj prilegali MVNblock in MVNorig podatki. Pri uvrščanju z L-SVM smo ob izboru spremenljivk glede na največjo F-statistiko ugotovili, da so vsi umetno generirani podatki, razen MVNblock podatkov, precenili realne vrednosti, pri čemer so se realnim rezultatom najbolj prilegali rezultati MVNorig podatkov, najslabše pa rezultati MVNblock in IND podatkov.

Pri uvrščanju realnih podatkov, ko je bila za zmanjšanje dimenzije podatkov uporabljena metoda PCA, opazimo pri uvrščanju z vsemi tremi metodami, da so razredi ob uporabi PCA slabše ločljivi med seboj, vpliv neravnotežja je večji kot pri rezultatih brez zmanjšanja dimenzije podatkov.

Pri uvrščanju umetno generiranih podatkov z metodo LDA ob uporabi PCA so rezultati MVNblock izrazito podcenili realne vrednosti, rezultati IND so jih izrazito precenili. Pri metodi uvrščanja L-SVM opazimo, da so tako rezultati IND kot tudi rezultati MVNblock podatkov realne rezultate izrazito podcenili. Pri metodi CART ob uporabi PCA za zmanjšanje dimenzije podatkov so izrazito odstopali od realnih vrednosti le IND podatki, ki so kazali veliko večji vpliv neravnotežja kot realni podatki. Rezultati MVNorig in ABS podatkov so se ob uporabi PCA pri vseh obravnavanih metodah uvrščanja dobro prilegali realnim podatkom.

Ob uporabi metode MDS pri uvrščanju realnih podatkov z metodo LDA smo opazili, da se je vpliv neravnotežja bistveno zmanjšal v primerjavi z rezultati brez uporabe MDS, kljub temu pa napovedne točnosti ob uporabi MDS v primeru treh različnih razredov na splošno niso bile višje od tistih brez uporabe MDS. Pri uvrščanju realnih podatkov z metodo CART ob uporabi MDS smo opazili, da metoda MDS ni odpravila vpliva neravnotežja, rezultati pa so bili podobni kot brez uporabe MDS. Pri uvrščanju realnih podatkov z metodo L-SVM smo ob uporabi MDS na splošno opazili slabše rezultate kot ob uporabi prilagojenega praga za uvrščanje. Uporaba MDS pri metodi L-SVM namreč ni odpravila vpliva neravnotežja.

Pri uvrščanju umetno generiranih podatkov ob uporabi metode MDS smo opazili, da so se rezultati IND in MVNblock podatkov v nekaterih primerih bolj približali realnim vrednostim kot brez uporabe MDS, vendar so rezultati pri IND in MVNblock na splošno še vedno izrazito odstopali od realnih rezultatov. Rezultati MVNorig in ABS podatkov pa so bili realnim rezultatom podobno blizu kot brez uporabe MDS. Pri uvrščanju z metodo CART smo opazili, da so bila odstopanja rezultatov umetno generiranih podatkov od realnih rezultatov na splošno manjša kot brez uporabe MDS.

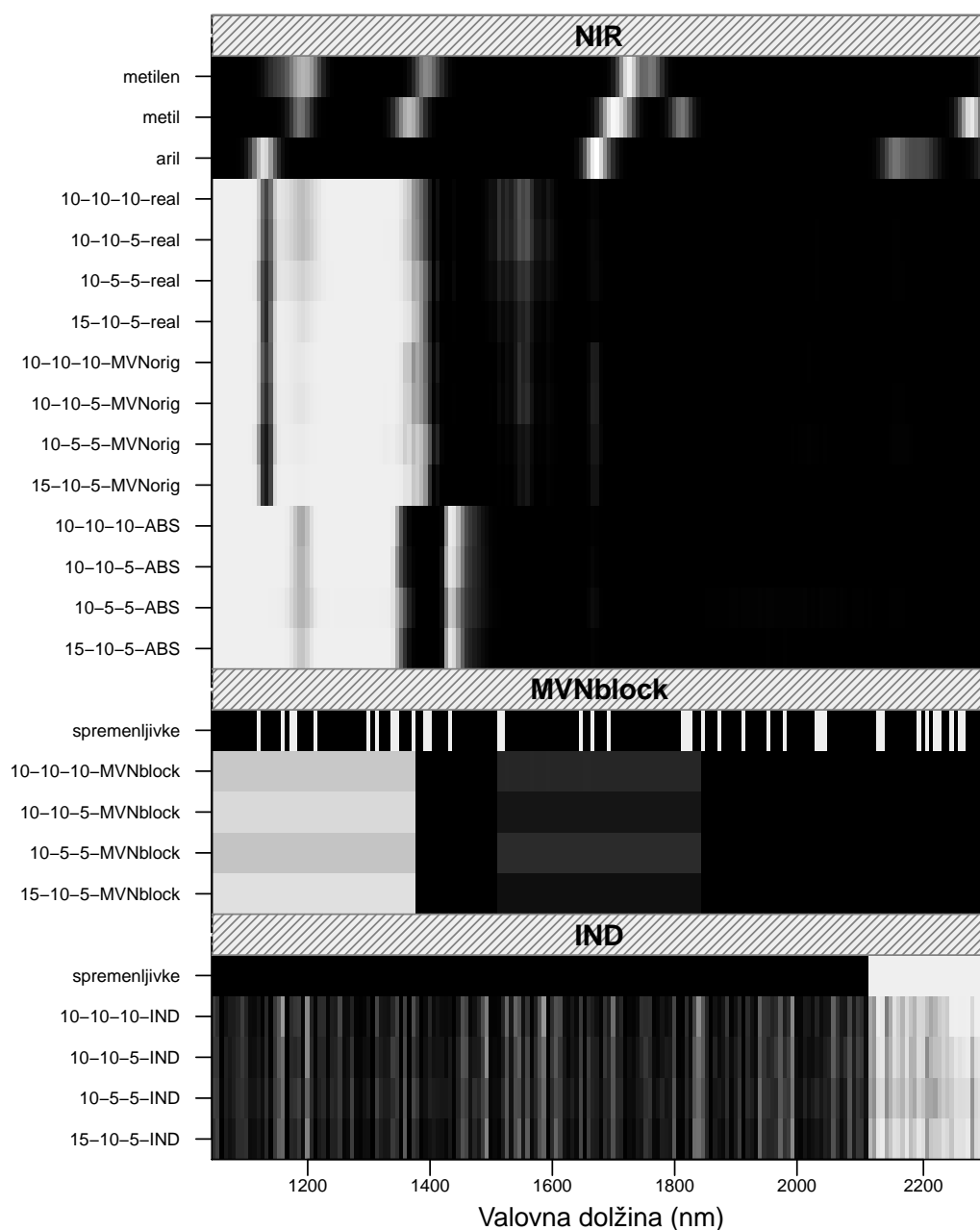
#### 4.3.3.3 Izbrane spremenljivke (Slike 4.3.16, 4.3.17, 4.3.18)

Napovedne točnosti posameznih razredov pri simulacijah uvrščanja v tri razrede (PS, PP in PE), pri katerih smo uporabili metode za zmanjšanje dimenzije podatkov (izbor 50 spremenljivk z največjo varianco, izbor 50 spremenljivk z največjo F-statistiko in PCA), so prikazane v Dodatku na Slikah D.3–D.5. V tem razdelku so prikazane spremenljivke, ki so bile v modele uvrščanja dejansko izbrane. Zanimalo nas je namreč, ali so bile v modele izbrane tiste spremenljivke, ki so bile med razredi dejansko različno izražene.

Na Slikah 4.3.16–4.3.18 so v vrsticah “metilen”, “metil” in “aril” prikazane absorpcije funkcionalnih skupin (bela – popolna absorpcija, črna – popoln odboj), ki smo jih dobili, kot je opisano v razdelku 3.4.4. Absorpcije v teh funkcionalnih skupinah so namreč vzrok za razlike v NIR spektrih razredov PS, PP in PE. V vrsticah “spremenljivke” so z belo barvo prikazane spremenljivke, ki so bile med razredi različno izražene pri MVNblock oz. IND podatkih. V ostalih vrsticah je za realne podatke (real) in za štiri vrste umetno generiranih podatkov (MVNorig, ABS, MVNblock in IND) in za različno stopnjo neravnotežja  $((n_{PS}, n_{PP}, n_{PE}) = (10, 10, 10), (10, 10, 5), (10, 5, 5), (15, 10, 5))$  prikazano, kolikokrat je bila posamezna spremenljivka izbrana pri izboru 50 spremenljivk z največjo varianco (Slika 4.3.16) ali največjo F-statistiko (Slika 4.3.17) oz. kakšna je bila povprečna utež na izbranih glavnih komponentah v primeru uporabe metode PCA za zmanjšanje dimenzije podatkov (Slika 4.3.18). Bela barva pomeni, da je bila spremenljivka vedno izbrana v model oz. je bila njena povprečna utež pri PCA enaka 0,25; črna pomeni, da spremenljivka ni bila nikoli izbrana v model oz. je bila njena povprečna utež pri PCA enaka 0.

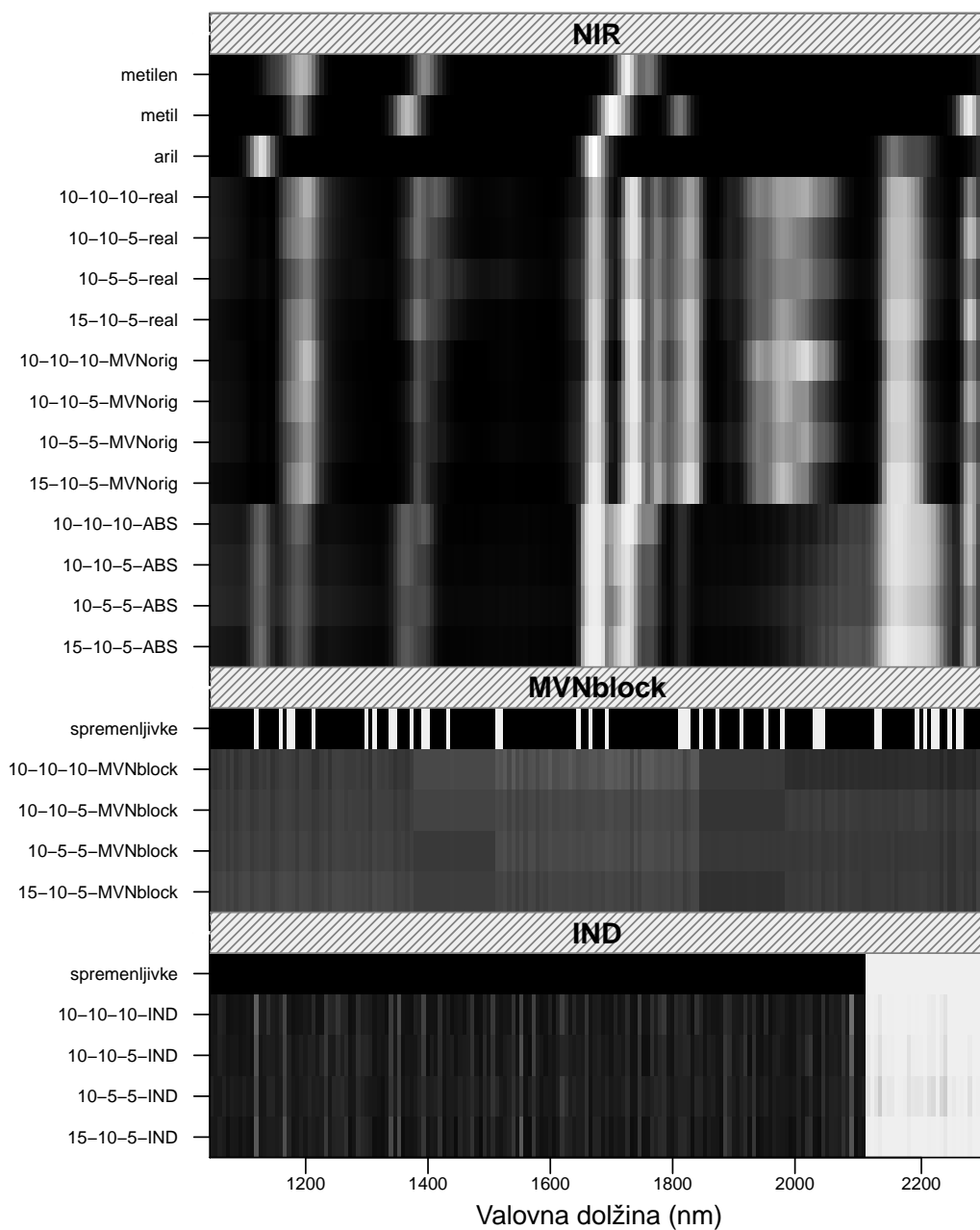
V primeru izbora spremenljivk glede na največjo varianco (Slika 4.3.16) opazimo, da pri uvrščanju MVNorig, ABS in realnih podatkov v modele niso bile izbrane spremenljivke, ki dejansko ločijo razrede, saj je bila večina spremenljivk izbranih iz začetka območja. Izbrane spremenljivke pri uvrščanju MVNblock podatkov nakazujejo bločno kovariančno matriko, na podlagi katere so bili ti podatki generirani. Pri IND podatkih je bila večina izbranih spremenljivk iz konca območja, kjer so dejansko bile spremenljivke, ki so bile med razredi različno izražene (razdelek 3.4.1). Neravnotežje na izbor spremenljivk najbolj vpliva pri MVNblock podatkih, pri ostalih podatkih pa ob spremembah velikosti razredov opazimo le rahle spremembe.

Spremenljivke, ki so bile izbrane na podlagi največje F-statistike (Slika 4.3.17), so se bolje ujemale s spremenljivkami, ki naj bi dejansko ločile razrede, kot tiste, ki so bile izbrane glede na največjo varianco. Pri podatkih ABS se je v primerjavi z drugimi vrstami podatkov izbor spremenljivk najboljše ujemal z absorpcijami metilena, metila in arila, kar je pričakovano, saj so bili podatki ABS generirani prav na podlagi informacije o teh absorpcijah. Izbor spremenljivk pri MVNorig in realnih podatkih se je med seboj zelo dobro ujemal, kar kaže na to, da so MVNorig podatki zelo podobni realnim. Izbrane spremenljivke pri MVNorig in realnih podatkih so se tudi sorazmerno dobro ujemale z absorpcijami opazovanih funkcionalnih skupin. Spremenljivke v območju med približno 1700 in 2100 nm so bile pri MVNorig in realnih podatkih pogosto izbrane v model, čeprav v tem pasu v literaturi nismo našli absorpcij skupin metila, metilena ali arila. Spremenljivke pri MVNblock so bile v primerjavi z izbranimi spremenljivkami na podlagi največje variance izbrane tukaj bolj enakomerno. Pri IND podatkih je bilo pri izbranih spremenljivkah glede na največjo F-statistiko v primerjavi z izbranimi spremenljivkami na podlagi največje variance še več spremenljivk, izbranih iz zadnjega dela, kjer so bile razlike med razredi dejansko največje.



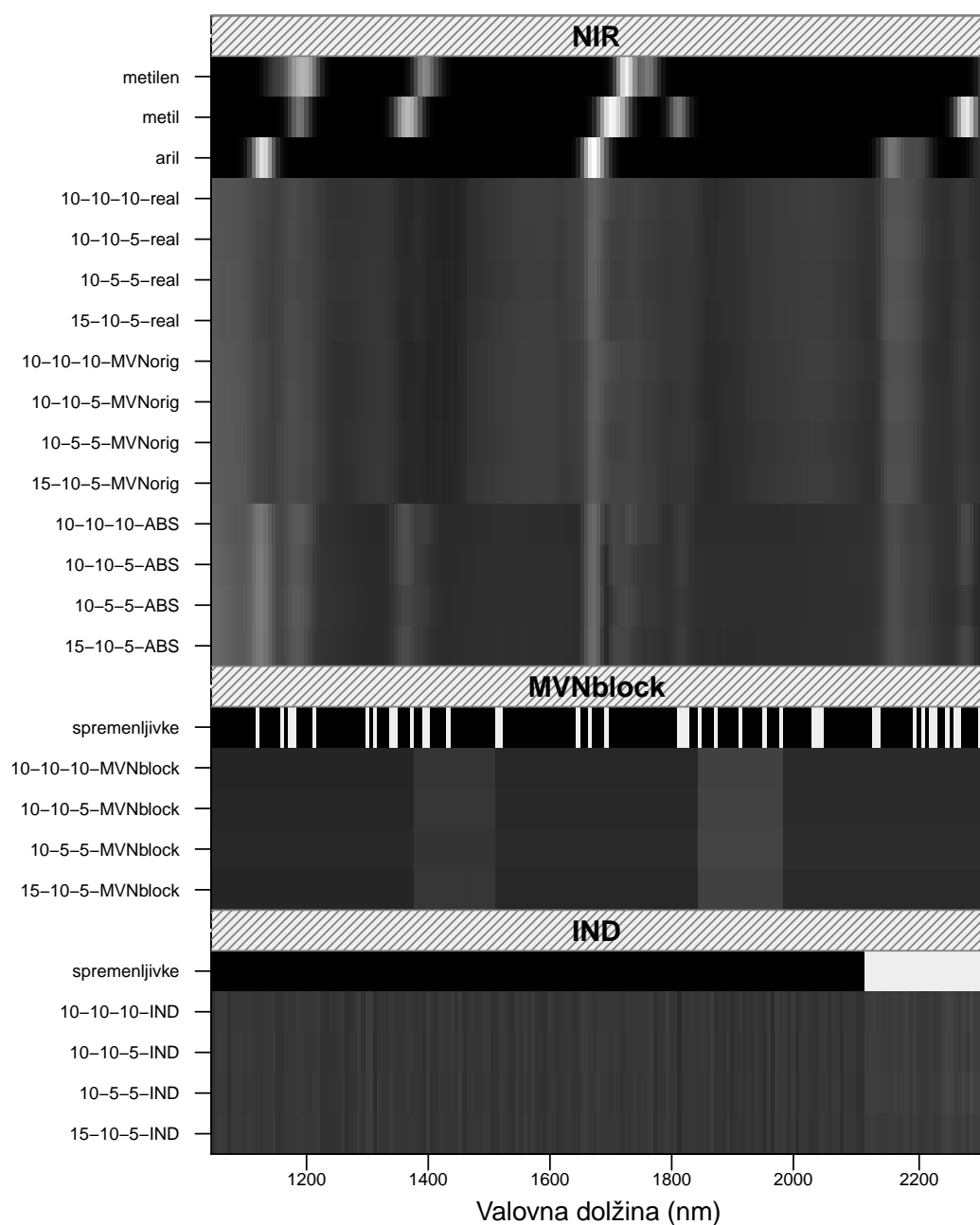
Slika 4.3.16: Izbrane spremenljivke pri uvrščanju v tri razrede (PS, PP in PE), pri čemer je bilo v model uvrščanja vključenih 50 spremenljivk z največjo varianco.

Doprinos posamezne spremenljivke v modele uvrščanja, pri katerih je bila za zmanjšanje dimenzije podatkov uporabljena metoda PCA, je prikazan na Sliki 4.3.18. V primerjavi z izbranimi spremenljivkami z največjo varianco ali največjo F-statistiko so bile tu spremenljivke v model "izbrane" enakomernejše, kar je glede na delovanje metode PCA pričakovano. Pri podatkih ABS ponovno (kot pri izboru z največjo F-statistiko) opazimo največjo skladnost z dejanskimi absorpcijami opazovanih funkcionalnih skupin. Tudi pri MVNorig in realnih podatkih se vidijo nekateri svetlejši pasovi, ki so skladni z absorpcijami opazovanih



Slika 4.3.17: Izbrane spremenljivke pri uvrščanju v tri razrede (PS, PP in PE), pri čemer je bilo v model uvrščanja vključenih 50 spremenljivk z največjo F-statistiko.

funkcionalnih skupin, vendar jih je manj in so manj izraziti kot pri ABS. Pri podatkih MVNblock opazimo močno povezanost med utežmi pri PCA in korelacijsko strukturo podatkov. Pri podatkih IND so bile uteži po spremenljivkah porazdeljene enakomerno.



Slika 4.3.18: Izbrane spremenljivke pri uvrščanju v tri razrede (PS, PP in PE), pri čemer so bile v model uvrščanja vključene glavne komponente, izračunane po metodi PCA, ki so pokrile 99 % celotne variabilnosti.



#### 4.3.4 Uvrščanje v 45 razredov

V Tabelah 4.3.3 in 4.3.4 so prikazane vrednosti mer A in G pri uvrščanju MVNorig podatkov v 45 razredov. Število enot in razlike med razredi smo poskusili generirati tako kot v realnih podatkih, kar je opisano v razdelku 3.4.3. Vrednosti mer A in G pri simuliranih podatkih so višje kot vrednosti, pridobljene pri uvrščanju realnih podatkov (Tabele 4.1.1, 4.1.2, 4.1.3 in 4.1.5), so pa med različnimi modeli uvrščanja podobno rangirane. Če bi se odločali za najboljšo metodo uvrščanja glede na rezultate pri simuliranih podatkih, bi se odločili za isto kombinacijo metod kot pri realnih podatkih.

Predobdelava	Zmanjšanje dim.	LDA	CART	L-SVM
brez	brez	0,85	0,11	0,56
	varianca	0,85	0,11	0,56
	F-statistika	0,90	0,24	0,79
	PCA	0,62	0,27	0,83
1. odvod	brez	0,98	0,38	0,94
	varianca	0,90	0,40	0,91
	F-statistika	0,90	0,38	0,91
	PCA	0,83	0,36	0,93
SNV	brez	<b>0,99</b>	<b>0,42</b>	0,93
	varianca	0,94	0,38	0,91
	F-statistika	0,92	0,40	0,93
	PCA	0,88	0,38	<b>0,95</b>

Tabela 4.3.3: **45 razredov (A-povprečje)**. Povprečne vrednosti A-povprečja iz 100 krat ponovljenega prečnega preverjanja z 10 pregibi pri uvrščanju podatkov simuliranih iz multivariatne normalne porazdelitve s parametri, ocenjenimi iz realnih podatkov (MVNorig). Število enot v posameznem razredu je bilo enako kot v realnih podatkih plastik. Najboljši rezultat za posamezno metodo uvrščanja (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	LDA	CART	L-SVM
brez	brez	0,85	0,09	0,50
	varianca	0,85	0,09	0,50
	F-statistika	0,90	0,19	0,77
	PCA	0,62	0,23	0,81
1. odvod	brez	0,98	0,35	<b>0,94</b>
	varianca	0,90	0,35	0,91
	F-statistika	0,90	0,35	0,91
	PCA	0,83	0,32	0,91
SNV	brez	<b>0,99</b>	<b>0,40</b>	0,90
	varianca	0,94	0,37	0,89
	F-statistika	0,92	0,39	0,91
	PCA	0,88	0,34	0,93

Tabela 4.3.4: **45 razredov (G-povprečje)**. Povprečne vrednosti mere G iz 100 krat ponovljenega prečnega preverjanja z 10 pregibi pri uvrščanju podatkov simuliranih iz multivariatne normalne porazdelitve s parametri, ocenjenimi iz realnih podatkov (MVNorig). Število enot v posameznem razredu je bilo enako kot v realnih podatkih plastik. Najboljši rezultat za posamezno metodo uvrščanja (najvišja vrednost v vsakem stolpcu) je zapisan krepko.



## Poglavje 5

# Razprava

Z različnimi kombinacijami metod predobdelave, zmanjšanja dimenzije podatkov, zmanjšanja vpliva neravnotežja ter metod uvrščanja smo zgradili 300 modelov za uvrščanje 45 vrst polimerov na osnovi podatkov bližnje infrardeče spektroskopije. Modele smo ovrednotili z različnimi merami in pristopi za vrednotenje uvrščanja ter izbrali najboljši model, ki napoveduje neznane enote z dovolj visoko stopnjo natančnosti, da je praktično uporaben za identificiranje polimernih materialov v muzejskih zbirkah. Tekom tega postopka smo zbirali podatke o delovanju 16 mer za vrednotenje uvrščanja, ki smo jih nato ovrednotili z analizo konkordance.

Metode uvrščanja in načrtovalne strategije pri uvrščanju smo proučili tudi s simulacijami, za katere je bilo potrebno najprej generirati umetne podatke. V ta namen smo predlagali nove metode generiranja podatkov. Nato smo v simulacijski študiji uvrščanja v dva, tri in 45 razredov primerjali rezultate na štiri različne načine pridobljenih umetno generiranih podatkov z realnimi podatki, da bi ugotovili, katera metoda je za generiranje NIRS podatkov najprimernejša.

### 5.1 Izbrani model za uvrščanje 45 vrst polimerov

Model za uvrščanje 45 vrst polimerov, ki je dosegel najvišje vrednosti mer za vrednotenje uvrščanja, je bil zgrajen na podatkih predobdelanih z metodo SNV in brez zmanjšanja dimenzije podatkov, uporabljena metoda uvrščanja pa je bila LDA. Vrednost A-povprečja, pridobljenega na podlagi 100 krat ponovljenega prečnega preverjanja z 10 pregibi je bila 0,88. Te vrednosti A-povprečja ni presegel tudi noben od obravnavanih modelov, pri katerih sta bili uporabljeni metodi MDS ali OAO.

Kot posledico neravnotežja smo v rezultatih opazili, da se napovedna točnost z večanjem števila enot v razredih zvišuje, medtem ko se variabilnost znižuje, kar je dobro znan pojav pri uvrščanju neuravnoteženih podatkov [29]. Kljub temu pa so bili nekateri zelo majhni razredi napovedani zelo dobro. Razlog za zelo dobro napovedno točnost majhnih razredov bi lahko bil v velikih razlikah v kemijski strukturi teh materialov od ostalih, iz literature je namreč znano tudi, da se vpliv neravnotežja z večanjem razlik med razredi zmanjšuje [29, 166].

Razlogov za napačne napovedi je več. Pogost vzrok napačnih napovedi je kemijska sorodnost med materiali, tako lahko npr. opazimo (Tabela 4.1.13), da so bile enote akrilonitril-butadien-

stirena (št. 5) v 39 % napovedane kot stiren-akrilonitril (št. 20) in obratno enote stiren-akrilonitrila (št. 20) so bile v 25 % napovedane kot akrilonitril-butadien-stiren (št. 5). Oba materiala sta namreč kopolimera stirena in akrilonitrila. Podobno sta celulozni-acetat-propionat (št. 19) in celulozni-acetat-butirat (št. 23) oba estra celuloznega acetata. Enote usnja (št. 18) so bile v 33 % napovedane kot kosti, zobje, roževina ... (št. 36), pri čemer so tako usnje kot materiali pod št. 36 pretežno sestavljeni iz beljakovin. Manjši delež (8 %) enot materialov pod št. 36, je bilo napovedanih tudi kot kazein formaldehid (št. 7), ki je umetna imitacija teh materialov. Podobno sta stiren butadien (št. 25) in poliizopren (št. 11) sintetična nadomestka naravne gume (št. 21). Kemijsko zelo podobni so si tudi materiali št. 6, 44 in 29 ter polikarbonat (št. 16) in polisulfon (št. 9). Včasih so vzrok napačnih napovedi slabe spektroskopske meritve. Pri nekaterih materialih je namreč težko pridobiti NIR spekter dobre kakovosti, ne da bi pri tem poškodovali vzorec [167]. Posebej če so vzorci stari in degradirani, kot je na primer v primeru materiala št. 28 (Polyurethane). V primeru celofana (št. 2) je kakovostno meritev težko pridobiti zaradi mehanskih lastnosti materiala. Celofan je namreč zelo tanek in popolnoma transparenten material. NIR meritve običajno dobimo tako, da material večkrat prepognemo. Pregibi materiala vplivajo na odboj svetlobe. Razlog za veliko število napačnih uvrstitev pri celofanu je tudi majhno število enot (3 enote). NIR spektri celofana se sicer izrazito razlikujejo od ostalih spektrov polimerov v naši podatkovni množici, vendar lahko na Sliki A.1 opazimo, da se meritve treh enot celofana iz naše podatkovne množice slabo prekrivajo. Napačne napovedi politetrafluoroetilena (št. 17) bi najverjetneje tudi lahko pripisali nekakovostni meritvi ene enote (od skupno šestih), ki jo lahko opazimo na Sliki A.2. Posamezne spektre, ki se močno razlikujejo od ostalih pri istem materialu, lahko opazimo tudi pri materialih polietilen tereftalat (št. 37), celulozni acetat (št. 24), stiren butadien (št. 25) ... (Slike A.1–A.3). Te enote verjetno tudi več prispevajo k napakam v napovedi. Nekaterih napačnih napovedi pa ne moremo razložiti s kemijsko zgradbo ali težavami v merilnem aparatu. Te vrste napake so verjetno posledica (ne)učinkovitosti modela uvrščanja in so iz zornega kota uporabnika resnično neželene.

## 5.2 Generiranje umetnih NIRS podatkov

Umetne podatke, generirane na štiri različne načine, smo na primeru simulacij uvrščanja neuravnoteženih podatkov primerjali z realnimi podatki. Pri prvem načinu so bili podatki generirani neodvisno iz normalne porazdelitve (IND); pri drugem načinu so bili generirani iz MVN porazdelitve z bločno kovariančno matriko (MVNblock); pri tretjem načinu so bili generirani iz MVN porazdelitve s parametri ocenjenimi iz realnih podatkov (MVNorig); pri četrtem načinu pa smo podatke generirali na podlagi teoretičnih absorpcij funkcionalnih skupin v NIR območju (ABS).

Na podlagi naših rezultatov se uvrščanju realnih podatkov najmanj približajo IND podatki, nekoliko bolj pa MVNblock podatki. Najbolj se realnemu stanju približajo MVNorig podatki. Pri tem načinu generiranja podatkov smo parametre MVN porazdelitve ocenili iz istih realnih podatkov, s katerimi smo kasneje primerjali rezultate uvrščanja. Podobnost rezultatov uvrščanja MVNorig in realnih podatkov bi bila zato lahko posledica pristranskosti. Zato smo v enem od obravnavanih primerov parametre MVN porazdelitve ocenili le na delu realnih podatkov (10 enot v vsakem razredu), ki smo jih nato izključili iz množice realnih podatkov, vključenih v simulacijsko študijo. S tem smo izključili pristranskost, saj so bile ocene parametrov MVN porazdelitve sedaj neodvisne od realnih podatkov, vključe-

nih v simulacijsko študijo. Izkazalo se je, da se na takšen način generirani (nepristranski) MVNorig podatki še vedno bolje približajo realnemu stanju, kot podatki generirani na druge tri načine. Podatki ABS se realnemu stanju večinoma tudi dobro približajo, čeprav nekoliko slabše kot podatki MVNorig. Vzrok za nekoliko slabše rezultate so najverjetneje zapletene absorpcije v NIR območju, zaradi katerih na podlagi kemijske strukture materiala ni mogoče natančno določiti absorpcijskih mest in njihove moči, kar je natančno razloženo v [60].

Naši rezultati kažejo, da način generiranja umetnih podatkov za namen raziskovanja statističnih metod s simulacijami zelo vpliva na rezultate uvrščanja. Simulacije z umetno generiranimi rezultati, ki se ne približajo dovolj lastnostim realnih podatkov, lahko privedejo do zavajajočih sklepov, na kar so opozarjali že raziskovalci na področju mikromrež [168, 169].

Predvsem na področju biologije in medicine je bila izražena kritika na način simuliranja podatkov, nezaupanje v metode preizkušene na podatkih, ki niso primerljivi z realnimi, in spodbuda k iskanju primernejših načinov testiranja metod in simuliranja podatkov [168, 169]. V zadnjem času lahko zato zasledimo več poskusov simuliranja podatkov, ki ohranijo strukturo realnih podatkov. Predstavljenih je bilo več algoritmov generiranja tako imenovanih *plasmode* podatkov, ki za osnovo vzamejo znano množico realnih podatkov, in nato simulirane podatke ustvarijo s kombinacijo prepletenega vzorčenja (angl. bootstrap) in dodajanja velikosti učinka obravnavanj, ki ga predhodno ocenijo iz realnih podatkov [170–172]. Tako generirani podatki ohranijo korelacijsko strukturo in variabilnost realnih podatkov, kljub temu pa je lastnost, ki je za raziskovalca pomembna, natančno znana.

Generiranje podatkov v naši študiji ne sledi nobenemu od predstavljenih algoritmov generiranja *plasmode* podatkov, saj se predstavljeni algoritmi, prilagojeni podatkom s področja mikromrež in analize preživetja, ki se bistveno razlikujejo od NIRS podatkov. Načini generiranja podatkov, predstavljeni v tem poglavju, zato niso tipični *plasmode* pristopi, se jim pa približajo v tem, da v postopku simulacije uporabijo nekatere karakteristike, ki so ocenjene iz realnih podatkov.

Za potrebe raziskovanja statističnih metod pri uvrščanju NIRS podatkov predlagamo, da se umetni podatki generirajo po metodi MVNorig pri pogoju, da je na voljo dovolj realnih podatkov. Če realnih podatkov ni dovolj in imamo informacije o absorpcijah v NIR območju, lahko uporabimo metodo ABS. Prednost metode ABS je tudi, da jo lahko uporabimo v primeru, ko realnih podatkov nimamo. V tem primeru lahko izberemo poljubne obstoječe funkcionalne skupine in njihove absorpcije, ali pa določimo funkcionalne skupine s fiktivno absorpcijo.

## 5.3 Predobdelave

Predobdelava je v analizi NIRS podatkov utečen postopek. V doktorskem delu smo obravnavali štiri metode predobdelave: brez predobdelave, kvantilna normalizacija, 1. odvod in SNV. Rezultati uvrščanja 45 vrst polimerov so pokazali, da predobdelava zelo vpliva na učinkovitost uvrščanja. Najslabše rezultate pri vseh metodah uvrščanja smo dobili, če nismo uporabili nobene predobdelave. S tem smo potrdili smiselnost uporabe predobdelave, kot utečenega postopka analize NIRS podatkov. Izbor najboljše metode predobdelave je sicer odvisen še od drugih metod uporabljenih v procesu izgradnje modela uvrščanja (me-

toda uvrščanja, metoda za zmanjšanje dimenzije podatkov, metoda za zmanjšanje vpliva neravnotežja).

Glede na naše rezultate pri uvrščanju realnih podatkov lahko trdimo, da je pri metodi LDA najuspešnejša predobdelava SNV, pri uvrščanju s CART 1. odvod, pri metodah k-NN kvantilna normalizacija, pri obeh obravnavanih metodah SVM pa sta predobdelavi SNV in kvantilna normalizacija primerljivo uspešni.

V simulacijski študiji smo testirali le predobdelavi SNV in 1. odvod. Izkazalo se je, da obe predobdelavi pozitivno vplivata na rezultate uvrščanja realnih podatkov in generiranih podatkov, ki so bili bolj podobni realnim (MVNorig in ABS). Simulacijska študija je potrdila ugotovitve, pridobljene pri uvrščanju 45 vrst polimerov, in sicer: pri uvrščanju z metodo LDA je uspešnejša predobdelava SNV, pri uvrščanju z metodo CART pa je uspešnejša predobdelava s 1. odvodom. Pri uvrščanju z metodo L-SVM težko rečemo, katera predobdelava je uspešnejša, saj so bili rezultati uvrščanja že brez uporabe predobdelave zelo dobri. Uporaba katerekoli od obeh predobdelav na splošno poslabšala uvrščanje IND podatkov.

Predobdelava podatkov je torej smiselna pri uvrščanju podatkov s podobnimi lastnostmi, kot jih imajo NIRS podatki. Pri uvrščanju neodvisnih podatkov, pa predobdelava rezultate večinoma poslabša. Pri uvrščanju z metodo LDA predlagamo predobdelavo SNV, pri uvrščanju z metodo CART pa predobdelavo s 1. odvodom. Pri uvrščanju z metodami SVM bi bile potrebne dodatne študije, s katerimi bi lahko potrdili prednost določene metode. Podobno bi bilo potrebno dodatno raziskati vpliv predobdelav na podatke s podobnimi lastnostmi, kot so jih imeli podatki MVNblock.

## 5.4 Zmanjšanje dimenzije podatkov in izbrane spremenljivke

Pri metodah za zmanjšanje dimenzije podatkov smo se osredotočili na izbor 50 spremenljivk na podlagi največje variance, izbor 50 spremenljivk na podlagi največje F-statistike in metodo PCA, kjer smo izmed vseh glavnih komponent v model uvrščanja izbrali toliko glavnih komponent, da so pokrile 99 % skupne variabilnosti.

Rezultati uvrščanja 45 vrst polimerov kažejo, da obravnavane metode zmanjšanja dimenzije podatkov negativno vplivajo na uvrščanje z metodama LDA in CART. Pri CART je izjema le metoda PCA v kombinaciji s predobdelavo s 1. odvodom. Pri metodi L-SVM je vpliv metod za zmanjšanje dimenzije podatkov odvisen od uporabljene predobdelave. Pri uvrščanju z R-SVM na uvrščanje pozitivno vpliva le metoda PCA, pri metodah k-NN pa izbor spremenljivk z največjo F-statistiko.

Rezultati simulacijske študije v večini primerov kažejo negativen vpliv uporabe metod za zmanjšanje dimenzije pri uvrščanju podatkov, ki imajo lastnosti, podobne NIRS podatkom (realni podatki, MVNorig in ABS). Izjema pri tem je metoda CART, pri kateri se je pokazalo rahlo izboljšanje rezultatov ob izboru spremenljivk na podlagi F-statistike. Izbor spremenljivk na podlagi največje variance in največje F-statistike na splošno izboljša rezultate uvrščanja podatkov IND in MVNblock. Pri uporabi metode PCA opazimo splošno poslabšanje rezultatov uvrščanja pri vseh vrstah podatkov. V večini obravnavanih primerov uvrščanja je uporaba PCA povečala vpliv neravnotežja. Iz rezultatov opazimo tudi izra-

zitejše poslabšanje napovedanih točnosti podobnejših razredov (razreda PE in PP). Kar kaže na to, da glavne komponente izračunane po metodi PCA, ki so vključene v modele uvrščanja, vsebujejo manjšo informacijo o razlikah med razredi kot začetne (originalne) spremenljivke. Z vključitvijo glavnih komponent v modele uvrščanja so torej razlike med razredi “navidezno” manjše. Manjše razlike med razredi so vzrok večjemu vplivu neravnotežja [29] in posledično slabšim rezultatom uvrščanja.

V [29] so ob izboru spremenljivk opazili povečan vpliv neravnotežja (povečana pristranskost v prid večjemu razredu), kar pri nas ni bilo izrazito, saj smo povečan vpliv neravnotežja opazili le pri metodi PCA, ki ni metoda izbora spremenljivk. Razlog za neskladje rezultatov bi lahko bile razlike v simulacijskih nastavitvah. V [29] lahko namreč opazimo, da se vpliv neravnotežja povečuje z večanjem števila originalnih spremenljivk, zato bi vzrok za povečanje neravnotežja lahko ležal v številu originalnih spremenljivk in ne toliko v samem izboru spremenljivk.

V študijah s področja mikromrež so se ukvarjali z izborom spremenljivk pri uvrščanju neuravnoteženih in visoko razsežnih podatkov. Blagus in Lusa [29] sta kljub nekoliko povečani pristranskosti v prid večjemu razredu ob izboru spremenljivk opazila višje vrednosti mer za vrednotenje uvrščanja pri vseh opazovanih metodah uvrščanja. Podobno sta opazili tudi Dudoit in Fridlyand [173] za metodi diagonalna LDA in k-NN, Lin in Chen [166] pa sta celo zaključila, da je izbor spremenljivk pri diagonalni LDA bistven tako v uravnoteženem kot v neuravnoteženem primeru. Rezultati teh študij, se razen v primeru uvrščanja z metodo k-NN ne skladajo z našimi rezultati uvrščanja polimerov v 45 razredov. Vzrok za neskladje rezultatov verjetno leži v različnih (korelacijskih) strukturah NIRS podatkov in podatkov mikromrež in različnem številu začetnih spremenljivk - pri mikromrežah je spremenljivk namreč nekaj 1000, pri NIRS podatkih pa nekaj 100. Rezultati simulacij so sicer pokazali pozitiven vpliv izbora spremenljivk pri IND podatkih in MVNblock podatkih, pri podatkih, ki imajo strukturo bolj podobno NIRS podatkom, pa ne.

Rezultati, ki kažejo, katere spremenljivke so bile v modele uvrščanja dejansko izbrane, razložijo nekatere rezultate uvrščanja ob uporabi metod za zmanjšanje dimenzije podatkov. Pri uporabi 50 spremenljivk z največjo varianco so bile le pri IND podatkih večinsko izbrane spremenljivke, ki so bile med razredi različno izražene, zato je ta metoda izboljšala uvrščanje pri podatkih IND. Zanimivo je, da je metoda izboljšala rezultate uvrščanja tudi pri podatkih MVNblock, čeprav pri teh podatkih večinoma niso bile izbrane med razredi različno izražene spremenljivke. Temu bi lahko bila vzrok visoka koreliranost, ki je povzročila, da so bile razlike med razredi razpoznavne tudi iz spremenljivk, ki sicer niso bile generirane kot med razredi različno izražene spremenljivke, so pa z njimi močno korelirale.

Pri izboru 50 spremenljivk z največjo F-statistiko je pri IND in MVNblock podatkih zgodba podobna kot pri izboru spremenljivk z največjo varianco. Pri ostalih podatkih (realni, MVNorig, ABS) je, z razliko od izbora spremenljivk z največjo varianco, sorazmerno velik delež izbranih spremenljivk ustrezal tistim, ki naj bi bile med razredi različno izražene. Rezultati uvrščanja so bili v nekaterih primerih sicer nekoliko boljši kot ob izboru spremenljivk z največjo varianco, niso pa bili boljši v primerjavi z uvrščanjem brez zmanjšanja dimenzije podatkov. Za uvrščanje NIRS podatkov torej ni dovolj upoštevati le nekaterih spremenljivk, saj rezultati kažejo, da bi bila lahko informacija o pripadnosti razredu porazdeljena po celotnem NIR območju.

Spremenljivke, ki so prispevale največji delež h glavnim komponentam, vključenim v modele uvrščanja, ko je bila za zmanjšanje dimenzije uporabljena metoda PCA, so bile skoraj

enakomerno porazdeljene po celi množici spremenljivk. Metoda PCA se torej ni izkazala učinkovito v iskanju med razredi različno izraženih spremenljivk.

## 5.5 Metode za zmanjšanje vpliva neravnotežja

Obravnavali smo tri metode za zmanjšanje vpliva neravnotežja: MDS, OAO in prilagojen prag za uvrščanje. Metoda MDS je različica metode asimetrični *bagging*. Pri osnovni metodi so enote v novo učno množico izbrane s ponavljanjem, pri MDS pa so izbrane brez ponavljanja. Za uporabo metode MDS smo se odločili na podlagi raziskave v [174], kjer so pri vzorčenju brez ponavljanja dobili boljše rezultate kot pri vzorčenju s ponavljanjem. Metodo OAO smo uporabili le pri uvrščanju realnih podatkov s 45 vrstami polimerov, ob tem je bila uporabljena vedno pri uvrščanju z metodo SVM v več razredov. Prilagojen prag za uvrščanje smo uporabili le pri uvrščanju z metodo SVM, kot je natančno opisano v razdelku 3.1.

Rezultati simulacijske študije so pokazali, da prilagojen prag v kombinaciji z OAO pri uvrščanju z metodo SVM izrazito zmanjša vpliv neravnotežja, pri metodi MDS pa v nekaterih primerih opazimo celo nižje vrednosti mere G pri visoki stopnji neravnotežja v primerjavi z uvrščanjem brez uporabe metod za zmanjšanje vpliva neravnotežja. Na uvrščanje s CART metoda MDS ne vpliva. Pri uvrščanju z LDA se, razen za IND in MVNblock podatke, metoda MDS ni izkazala za uspešno pri uvrščanju v dva razreda, uspešno pa je zmanjšala vpliv neravnotežja pri uvrščanju v tri razrede. Razlog za to, da je bila metoda uspešna pri uvrščanju v tri razrede, pri uvrščanju v dva razreda pa ne, je najverjetneje v tem, da pri uvrščanju v tri razrede nismo obravnavali tako visoke stopnje neravnotežja kot pri uvrščanju v dva razreda. Pri uvrščanju v dva razreda so se ob uporabi MDS izrazito zmanjšale napovedne točnosti samo v primeru največje stopnje neravnotežja ( $k_1 = 0,1$  ali  $k_1 = 0,9$ ). Pri tako visoki stopnji neravnotežja so bile velikosti manjšega razreda zelo majhne. Metoda MDS pa deluje tako, da ustvari novo učno množico, v kateri je velikost vseh razredov enaka velikosti najmanjšega razreda. Učna množica je bila torej v tem primeru zelo majhna. Vzrok za slabo napoved uvrščanja pri visoki stopnji neravnotežja ob uporabi MDS torej ni v vplivu neravnotežja, ampak v majhni velikosti vzorca (angl. small sample bias).

Pri uvrščanju 45 vrst polimerov smo ob uporabi MDS dobili podobne rezultate kot v simulacijski študiji: pri LDA so se mere za vrednotenje uvrščanja rahlo povečale, pri CART v večini primerov metoda MDS ni vplivala, pri uvrščanju z R-SVM in L-SVM pa je bil pozitiven oz. negativen vpliv odvisen še od uporabljenih metod predobdelave in zmanjšanja dimenzije podatkov. Kljub temu da so bile vrednosti mer za vrednotenje uvrščanja pri LDA na splošno nekoliko višje ob uporabi MDS kot brez uporabe MDS, metoda MDS ni izboljšala rezultatov uvrščanja pri najboljšem izbranem modelu zgrajenem s SNV predobdelavo, brez zmanjšanja dimenzije podatkov in metodo uvrščanja LDA. Mere za vrednotenje uvrščanja pri tej kombinaciji metod kažejo, da sta modela z uporabo MDS in brez uporabe MDS približno enako učinkovita, zato metode MDS nismo vključili v končni izbrani model.

Pri uvrščanju 45 vrst polimerov ob uporabi metode OAO so se pri uvrščanju z LDA rezultati v primerjavi z rezultati brez uporabe metod za zmanjšanje vpliva neravnotežja izboljšali le v kombinaciji z zmanjšanjem dimenzije z metodo PCA. Metoda OAO se je izkazala za zelo učinkovito pri uvrščanju z metodo CART in R-SVM. Izboljšanje rezultatov uvrščanja z OAO pristopom pri metodi R-SVM je nenavadno, saj je za uvrščanje v več razredov metoda



OAo že vgrajena v uporabljeno funkcijo *svm* iz paketa *e1071* [147]. Sklepamo lahko le, da je bil naš algoritem za metodo OAo učinkovitejši od algoritma, ki je implementiran v funkciji *svm*, vendar tega ne moremo potrditi, ker nimamo dostopa do kode funkcije *svm*.

V simulacijski študiji smo pri nekaterih primerih uvrščanja z metodo L-SVM opazili, da so bili večji razredi slabše napovedani kot manjši (npr. Slika 4.3.11). Vzrok za to je najverjetneje v “preveč” prilagojenem pragu, ki smo ga uporabili kot del metode L-SVM. Lin in Chen [166] sta opozorila, da lahko prilagoditev praga, ki je odvisna samo od velikosti razredov, povzroči preprilagojenost v primeru, ko sta stopnja neravnotežja ali velikost učne množice veliki.

## 5.6 Metode uvrščanja

Metode uvrščanja, ki smo jih v doktorskem delu natančneje preučili, so bile: LDA, CART, SVM z radialno jedrno funkcijo (R-SVM), SVM z linearno jedrno funkcijo (L-SVM) in metode k-NN z 1, 3 in 5 sosedi.

Pri uvrščanju 45 vrst polimerov smo najboljše rezultate dobili z metodo LDA, sledila ji je metoda L-SVM, pri kateri smo uporabili tudi glede na velikosti razredov prilagojen prag za uvrščanje. Podobne rezultate kot z metodo L-SVM smo dobili z metodo 1-NN, slabše delovanje pa smo po vrsti opazili pri 3-NN, 5-NN in CART.

Vzrok, da metoda k-NN z manjšim številom sosedov deluje bolje, kot če število sosedov povečamo, je v majhnem številu enot v razredih, kar sta trdili tudi Dudoit in Fridlyand [173]. Pri podatkih polimerov so imeli najmanjši razredi velikost 3, kar pomeni, da sta lahko bili pri postopku prečnega preverjanja v učni množici le dve enoti iz posameznega razreda. Napovedna točnost takšnih razredov je bila pri uvrščanju z metodama 3-NN ali 5-NN seveda enaka 0.

V simulacijski študiji smo proučevali le metode LDA, CART, R-SVM in L-SVM. Delovanje metod SVM smo preizkusili tudi ob uporabi uteževanja razredov in prilagojenega praga za uvrščanje. Metoda SVM se je izkazala najuspešnejša ob uporabi linearne jedrne funkcije in prilagojenega praga za uvrščanje (L-SVM). Prilagoditev praga se je izkazala za uspešnejši pristop za obravnavo neravnotežja kot uteževanje razredov tudi pri R-SVM. Deloma bi bila lahko vzrok metoda OAo, ki smo jo uporabili v kombinaciji s prilagojenim pragom. Pri uvrščanju 45 razredov plastik smo namreč opazili, da uporaba metode OAo izboljša rezultate uvrščanja z metodo R-SVM, kar je sicer nekoliko nenavadno, saj je metoda OAo že implementirana v funkciji *svm*, kar smo komentirali že v razdelku 5.5.

V simulacijski študiji se je pri uvrščanju v dva in tri razrede najuspešnejša izkazala metoda L-SVM, sledila je LDA in nato CART. Slabo učinkovitost metode CART bi lahko pripisali veliki občutljivosti na neravnotežje, na katero so opozorili že v [29] pri proučevanju uvrščanja z metodo naključnih gozdov, ki deluje tako, da združuje več različnih modelov CART. Zanimivo je, da se je v simulacijski študiji pri uvrščanju v dva in tri razrede tudi pri realnih podatkih za uspešnejšo izkazala metoda L-SVM, pri uvrščanju polimerov v 45 razredov pa je bila uspešnejša metoda LDA. Uvrščanje 45 vrst polimerov je bil že iz vidika števila razredov zapletenejši problem, kot problemi uvrščanja v dva ali tri razrede v simulacijski študiji. Slabše delovanje metode L-SVM v kompleksnejši okoliščini bi morda lahko pripisali večjemu številu razredov in preveč prilagojenemu pragu za uvrščanje, ki smo ga komentirali

že v 5.5. To deloma potrdijo rezultati, ki smo jih dobili s simulacijo uvrščanja v 45 razredov na podatkih MVNorig, kjer je bil najuspešnejši model zgrajen z isto kombinacijo metod kot pri uvrščanju realnih podatkov 45 vrst polimerov, pri kateri je bila metoda uvrščanja LDA. Vendar bi morali slabše delovanje metode L-SVM v kompleksnejših okoliščinah bolj sistematično raziskati.

Naša študija je ponovno potrdila nenavadno dobro delovanje metode LDA kljub neuravnoteženi učni množici in kršeni predpostavki o neodvisnosti spremenljivk, kar sta prvič izpostavili Dudoit in Fridlyand [173], kasneje pa tudi študije [45] in [166]. Tudi metodi CART in L-SVM sta kljub prisotnosti koreliranih spremenljivk dobro delovali, saj so rezultati neodvisnih podatkov IND v simulacijski študiji pogosto izkazali slabšo napovedno točnost v primerjavi z ostalimi vrstami podatkov, ki so imeli korelirane spremenljivke. Slabše delovanje metode CART v primerjavi z metodama LDA in L-SVM, ki je bilo še posebej očitno pri uvrščanju realnih podatkov v 45 razredov polimerov, pa bi glede na rezultate simulacij lažje pripisali vplivu neravnotežja kot koreliranim spremenljivkam. V literaturi je sicer pri algoritmu naključnih dreves izpostavljena pristranskost, ki je posledica koreliranih spremenljivk, kot rešitev pa so predlagane različne mere, ki merijo pomembnost spremenljivk, ali različni algoritmi za izbor spremenljivk [18, 19]. Kot uspešno rešitev pri odpravljanju pristranskosti zaradi koreliranosti spremenljivk se je v literaturi izkazal izbor spremenljivk po skupinah, pridobljenih z metodo hierarhičnega razvrščanja spremenljivk v kombinaciji z regresijskimi metodami s kaznijo: kot sta regresijski metodi *lasso* in *ridge* [19, 175].

## 5.7 Skladnost mer za vrednotenje uvrščanja

Številne mere za vrednotenje uvrščanja so se razvile zaradi slabega delovanja nekaterih mer, kot je npr. skupna napovedna točnost pri uvrščanju neuravnoteženih podatkov. V doktorskem delu smo s 16 merami ovrednotili uvrščanje različnih klasifikacijskih modelov pri uvrščanju 45 vrst polimerov. V večini obravnavanih primerov so mere, kljub njihovi različni naravi, skladno izbrale isti najboljši model. Ob tem kažejo rezultati visoko stopnjo usklajenosti na splošno (ne le pri najboljšem modelu) z vrednostjo Kendallovega koeficienta konkordance  $W = 0,95$ . Slabšo usklajenost mer smo opazili pri modelih uvrščanja, ki so na splošno manj učinkoviti, kot so bili v našem primeru modeli zgrajeni z metodo CART.

Rezultati pri meri G se najslabše skladajo z ostalimi merami. Razlog za to je, da je vrednost mere G enaka 0, če je napovedna točnost vsaj enega od razredov enaka 0. Mera G torej zavzame najnižjo vrednost tako v primeru, ko je napovedna točnost le enega razreda enaka 0, kot tudi v primeru, ko je napovedna točnost vseh razredov enaka 0 (vse enote napovedane napačno). V primeru velikega števila razredov je med tema dvema možnostma veliko možnih izidov, med katerimi mera G ne razlikuje, druge mere pa. Pri meri G-001, ki poskuša odpraviti ta problem, opazimo večjo usklajenost z drugimi merami, čeprav zaradi podobnih lastnosti kot mera G od drugih mer še vedno odstopa. V nadaljnjem delu bi bilo smiselno preučiti, ali bi lahko s primernejšim popravkom mere G dosegli večjo skladnost z drugimi merami za vrednotenje uvrščanja. Največjo usklajenost smo opazili med merami PA, K in MCC ter med merama H in RCI.

Naši rezultati so v nasprotju z rezultati v študiji [176], kjer sta bili meri K in MCC manj usklajeni z mero PA kot mera Pacc. Podobno kot v naši študiji so rezultati v [104] pokazali

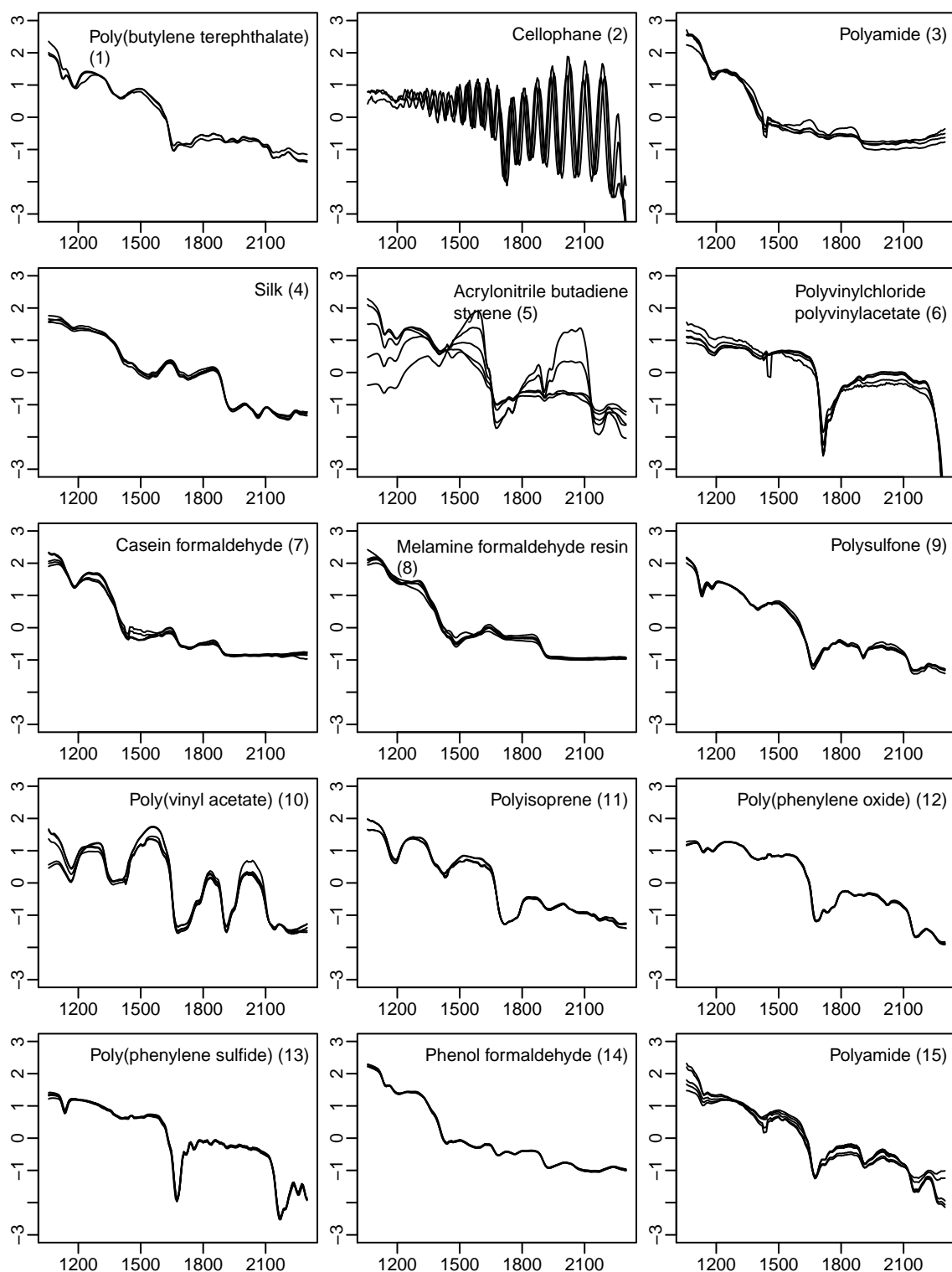
visoko usklajenost mer PA in K, v študiji [138] pa so pokazali podobnost med merama PA in MCC. Naši rezultati so tudi v skladu z opažanji v [113], kjer so prvič predstavili mero CEN in jo primerjali z merama PA in RCI, ter zaključili, da je mera CEN natančnejša, a usklajena z merama PA in RCI. Tudi pri nas je bila namreč mera CEN sorazmerno usklajena s tema merama, čeprav Kendallova korelacijska koeficienta pri parih PA in CEN ter RCI in CEN nista izstopala od ostalih.

Avtorji v različnih študijah (glej razdelek 2.2.5) opozarjajo na občutljivost mer za vrednotenje uvrščanja na neravnotežje. Ta občutljivost je verjetno najbolj znana pri meri PA. Glede na usklajeno delovanje vseh obravnavanih mer lahko sklepamo, da so na neravnotežje občutljive vse mere, še posebej pa meri K in MCC, ki sta z mero PA najbolj usklajeni, kot so podobno sklepali v [138]. V nadaljevanju bi bilo dobro preveriti še delovanje mer na podlagi AUC, ki naj bi bile na neravnotežje manj občutljive [92, 104], v ta namen pa bi bilo najprej potrebno najti učinkovito posplošitev mere AUC za uvrščanje v več razredov.

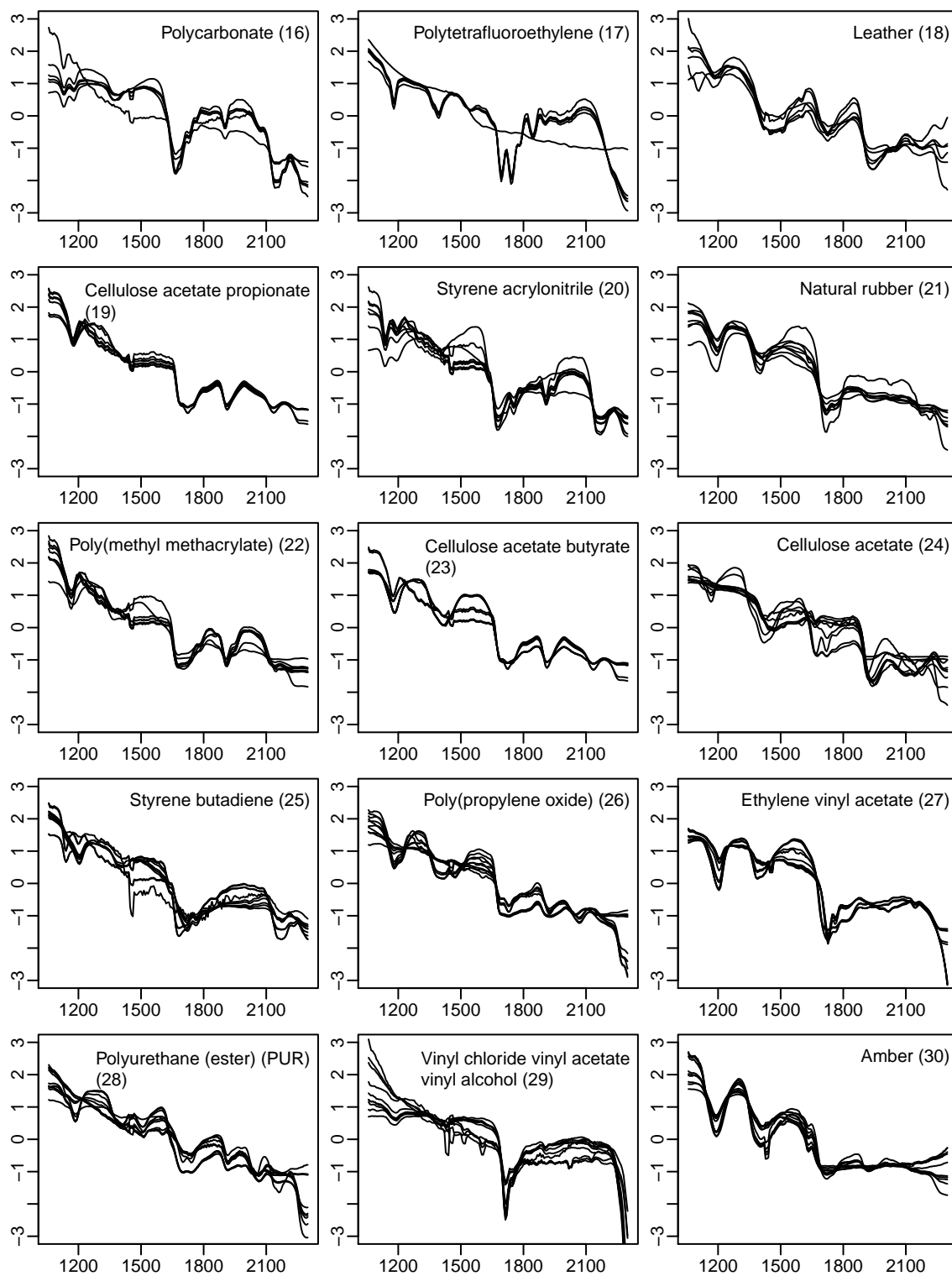


Dodatek A

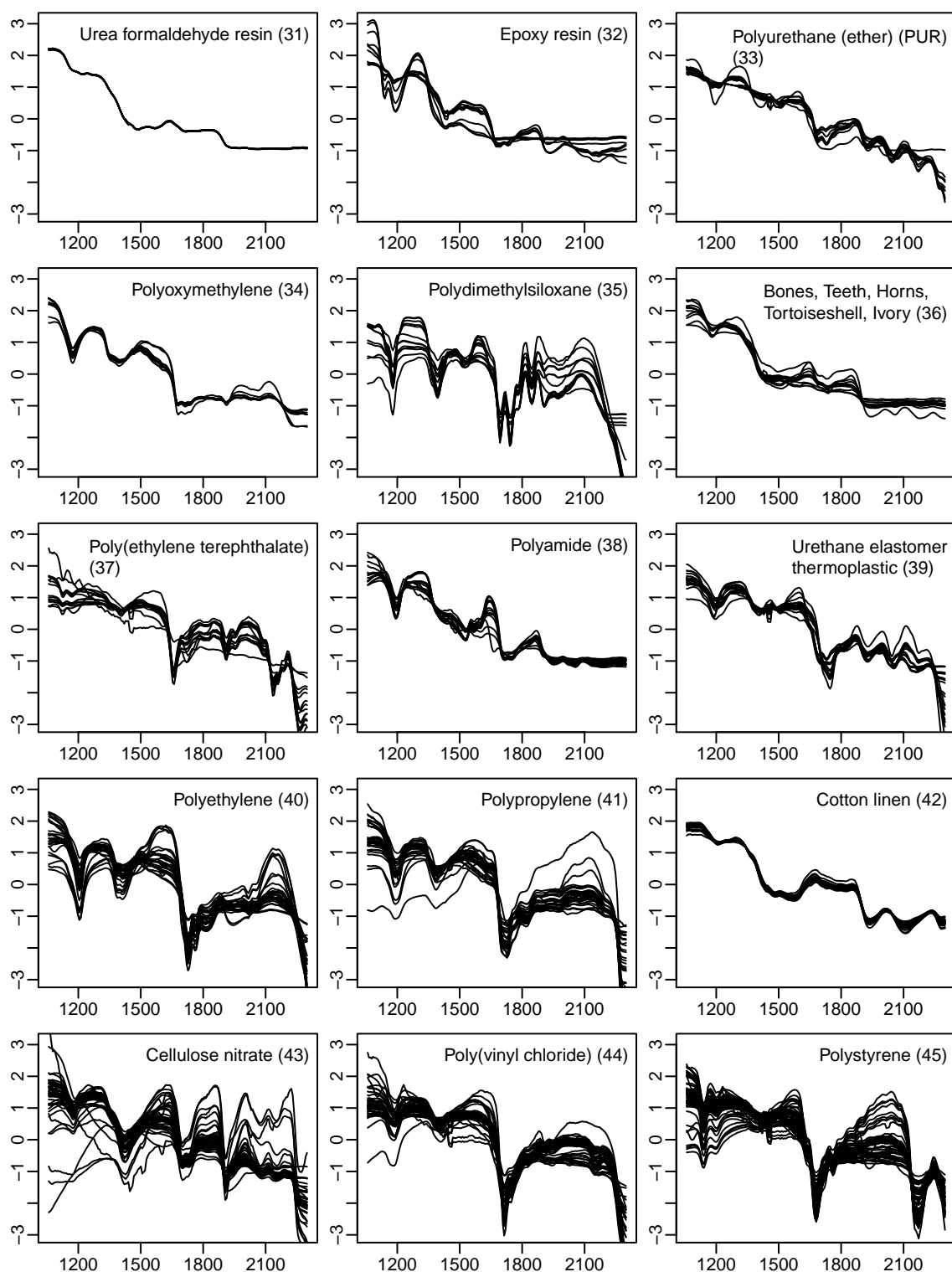
Slike spektrov polimerov



Slika A.1: Spektri polimerov z oznakami 1–15 po uporabi SNV predobdelave. Oznake so enake kot v Tabeli 3.1.1



Slika A.2: Spektri polimerov z oznakami 16–30 po uporabi SNV predobdelave. Oznake so enake kot v Tabeli 3.1.1



Slika A.3: Spektri polimerov z oznakami 31–45 po uporabi SNV predobdelave. Oznake so enake kot v Tabeli 3.1.1



## Dodatek B

Rezultati uvrščanja plastik z  
metodo najbližjih sosedov z več  
kot enim sosedom

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,53	0,48	0,00	0,23	0,51	0,50	0,47	0,51	0,56	0,42	0,35	-0,02	0,71	0,61	0,63	0,61
	varianca	0,32	0,28	0,00	0,06	0,28	0,29	0,27	0,29	0,40	0,24	0,19	-0,42	0,56	0,45	0,47	0,45
	F-statistika	0,58	0,51	0,00	0,27	0,57	0,53	0,50	0,57	0,58	0,45	0,38	0,04	0,75	0,65	0,67	0,65
	PCA	0,49	0,44	0,00	0,20	0,46	0,44	0,42	0,46	0,52	0,38	0,31	-0,13	0,69	0,59	0,60	0,59
kvant. norm.	brez	0,81	0,74	0,00	0,49	0,80	0,77	0,74	0,80	0,78	0,69	0,64	0,51	0,89	0,83	0,84	0,83
	varianca	0,78	0,70	0,00	0,46	0,76	0,75	0,70	0,76	0,75	0,64	0,59	0,45	0,87	0,80	0,82	0,80
	F-statistika	<b>0,83</b>	<b>0,78</b>	<b>0,72</b>	<b>0,68</b>	<b>0,82</b>	<b>0,82</b>	<b>0,77</b>	<b>0,82</b>	<b>0,81</b>	<b>0,71</b>	<b>0,66</b>	<b>0,60</b>	<b>0,90</b>	<b>0,85</b>	<b>0,86</b>	<b>0,85</b>
	PCA	0,77	0,70	0,00	0,55	0,76	0,75	0,70	0,76	0,75	0,64	0,58	0,45	0,86	0,80	0,81	0,80
1. odvod	brez	0,80	0,71	0,00	0,49	0,79	0,73	0,71	0,79	0,75	0,66	0,60	0,44	0,89	0,84	0,85	0,84
	varianca	0,78	0,69	0,00	0,46	0,77	0,72	0,68	0,77	0,74	0,63	0,57	0,41	0,87	0,82	0,82	0,82
	F-statistika	0,81	0,73	0,36	0,62	0,80	0,75	0,72	0,80	0,76	0,67	0,60	0,48	0,89	<b>0,85</b>	0,85	<b>0,85</b>
	PCA	0,80	0,70	0,00	0,48	0,79	0,72	0,70	0,79	0,74	0,65	0,59	0,42	0,88	0,84	0,84	0,84
SNV	brez	0,78	0,72	0,00	0,54	0,77	0,76	0,72	0,77	0,76	0,65	0,60	0,47	0,87	0,81	0,83	0,81
	varianca	0,77	0,69	0,00	0,45	0,76	0,71	0,68	0,76	0,73	0,62	0,56	0,40	0,86	0,80	0,82	0,80
	F-statistika	0,81	0,76	0,00	0,60	0,80	0,79	0,75	0,80	0,79	0,69	0,64	0,55	0,88	0,84	0,84	0,84
	PCA	0,78	0,71	0,00	0,56	0,76	0,76	0,71	0,76	0,76	0,64	0,59	0,47	0,87	0,81	0,82	0,81

Tabela B.1: **3-NN (10-CV)** Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo 3-NN. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,48	0,42	0,00	0,19	0,45	0,42	0,40	0,45	0,50	0,34	0,28	-0,16	0,68	0,57	0,58	0,57
	varianca	0,30	0,28	0,00	0,05	0,27	0,24	0,24	0,27	0,38	0,20	0,16	-0,48	0,56	0,44	0,46	0,44
	F-statistika	0,55	0,46	0,00	0,19	0,53	0,47	0,45	0,53	0,55	0,39	0,33	-0,07	0,73	0,62	0,64	0,62
	PCA	0,45	0,40	0,00	0,18	0,42	0,38	0,37	0,43	0,48	0,32	0,25	-0,22	0,66	0,56	0,57	0,56
kvant. norm.	brez	0,78	0,70	0,00	0,47	0,77	<b>0,74</b>	0,70	0,77	0,75	0,63	0,58	0,44	0,87	0,81	0,83	0,81
	varianca	0,73	0,65	0,00	0,35	0,72	0,69	0,64	0,72	0,71	0,59	0,54	0,34	0,85	0,78	0,80	0,78
	F-statistika	<b>0,81</b>	<b>0,74</b>	0,00	<b>0,49</b>	<b>0,80</b>	<b>0,74</b>	<b>0,72</b>	<b>0,80</b>	<b>0,77</b>	<b>0,66</b>	<b>0,61</b>	<b>0,48</b>	<b>0,89</b>	<b>0,83</b>	<b>0,84</b>	<b>0,83</b>
	PCA	0,74	0,66	0,00	0,36	0,73	0,66	0,64	0,73	0,70	0,57	0,52	0,32	0,85	0,79	0,80	0,79
1. odvod	brez	0,75	0,64	0,00	0,37	0,74	0,64	0,62	0,74	0,68	0,56	0,51	0,28	0,86	0,80	0,82	0,80
	varianca	0,72	0,61	0,00	0,30	0,70	0,59	0,59	0,70	0,65	0,53	0,48	0,20	0,84	0,78	0,79	0,78
	F-statistika	0,76	0,64	0,00	0,39	0,75	0,65	0,62	0,75	0,68	0,56	0,51	0,29	0,87	0,82	0,83	0,82
	PCA	0,75	0,64	0,00	0,36	0,74	0,63	0,62	0,74	0,68	0,56	0,51	0,27	0,86	0,80	0,81	0,80
SNV	brez	0,77	0,67	0,00	0,37	0,76	0,68	0,66	0,76	0,72	0,60	0,55	0,36	0,86	0,80	0,82	0,80
	varianca	0,73	0,62	0,00	0,32	0,72	0,63	0,61	0,72	0,68	0,55	0,49	0,26	0,84	0,77	0,79	0,77
	F-statistika	0,78	0,70	0,00	0,42	0,77	0,71	0,68	0,77	0,74	0,62	0,57	0,41	0,87	0,82	0,83	0,82
	PCA	0,75	0,66	0,00	0,36	0,74	0,67	0,65	0,74	0,71	0,58	0,54	0,33	0,86	0,79	0,81	0,79

Tabela B.2: **5-NN (10-CV)** Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 100 krat ponovljenim prečnim preverjanjem z 10 pregibi pri uvrščanju z metodo 5-NN. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,46	0,41	0,00	0,12	0,44	0,44	0,40	0,44	0,52	0,34	0,30	-0,15	0,75	0,65	0,68	0,65
	varianca	0,28	0,25	0,00	0,04	0,25	0,25	0,23	0,25	0,39	0,20	0,17	-0,50	0,64	0,54	0,58	0,54
	F-statistika	0,52	0,45	0,00	0,14	0,50	0,48	0,43	0,50	0,54	0,37	0,34	-0,07	0,78	0,68	0,72	0,68
	PCA	0,43	0,38	0,00	0,10	0,40	0,39	0,36	0,41	0,48	0,31	0,27	-0,23	0,74	0,64	0,67	0,64
kvant. norm.	brez	0,77	0,71	0,00	0,37	0,76	0,73	0,69	0,76	0,76	0,63	0,61	0,44	0,90	0,84	0,86	0,84
	varianca	0,73	0,65	0,00	0,30	0,72	0,68	0,64	0,72	0,72	0,58	0,55	0,33	0,88	0,82	0,84	0,82
	F-statistika	<b>0,80</b>	<b>0,74</b>	0,00	<b>0,41</b>	<b>0,79</b>	<b>0,75</b>	<b>0,72</b>	<b>0,79</b>	<b>0,78</b>	<b>0,66</b>	<b>0,64</b>	<b>0,49</b>	<b>0,91</b>	<b>0,87</b>	<b>0,88</b>	<b>0,87</b>
	PCA	0,77	0,70	0,00	0,37	0,76	0,73	0,69	0,76	0,75	0,63	0,60	0,43	0,90	0,84	0,86	0,84
1. odvod	brez	0,74	0,65	0,00	0,32	0,73	0,68	0,64	0,73	0,70	0,58	0,55	0,33	0,90	0,84	0,86	0,84
	varianca	0,71	0,62	0,00	0,27	0,70	0,64	0,60	0,70	0,68	0,54	0,51	0,25	0,88	0,82	0,84	0,82
	F-statistika	0,75	0,66	0,00	0,33	0,74	0,68	0,64	0,74	0,70	0,58	0,55	0,33	0,90	0,85	0,87	0,85
	PCA	0,74	0,65	0,00	0,31	0,73	0,66	0,63	0,73	0,70	0,57	0,54	0,31	0,90	0,84	0,86	0,84
SNV	brez	0,74	0,67	0,00	0,33	0,73	0,69	0,65	0,73	0,72	0,59	0,56	0,36	0,89	0,83	0,85	0,83
	varianca	0,70	0,60	0,00	0,26	0,68	0,62	0,59	0,69	0,67	0,52	0,49	0,23	0,87	0,80	0,83	0,80
	F-statistika	0,77	0,72	0,00	0,40	0,76	0,73	0,70	0,76	0,76	0,63	0,61	0,45	0,90	0,85	0,86	0,85
	PCA	0,73	0,66	0,00	0,33	0,72	0,68	0,65	0,72	0,72	0,58	0,55	0,35	0,88	0,82	0,84	0,82

Tabela B.3: **3-NN (10-CV)** Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo 3-NN. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.

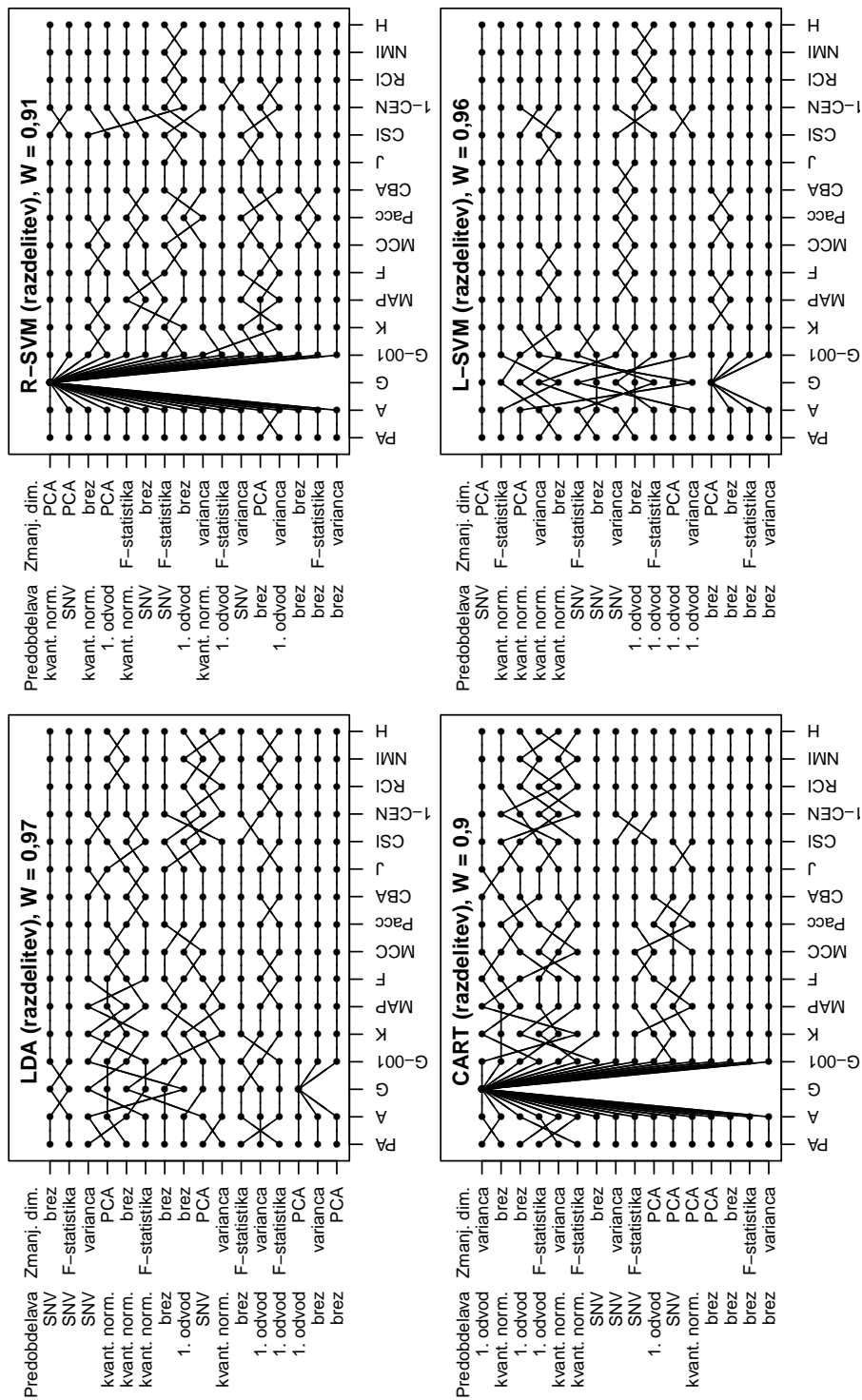
Predobdelava	Zmanjšanje dim.	PA	A	G	G-001	K	MAP	F	MCC	Pacc	CBA	J	CSI	1-CEN	RCI	NMI	H
brez	brez	0,42	0,35	0,00	0,08	0,40	0,35	0,32	0,40	0,46	0,27	0,23	-0,30	0,73	0,63	0,66	0,63
	varianca	0,28	0,24	0,00	0,03	0,24	0,22	0,21	0,24	0,39	0,18	0,15	-0,53	0,64	0,52	0,56	0,52
	F-statistika	0,48	0,40	0,00	0,10	0,46	0,41	0,37	0,46	0,51	0,32	0,28	-0,19	0,76	0,65	0,69	0,65
	PCA	0,40	0,34	0,00	0,07	0,38	0,33	0,31	0,38	0,45	0,26	0,22	-0,33	0,72	0,62	0,65	0,62
kvant. norm.	brez	0,71	0,61	0,00	0,22	0,70	0,61	0,58	0,70	0,68	0,52	0,50	0,22	0,88	0,80	0,84	0,80
	varianca	0,67	0,56	0,00	0,19	0,66	0,59	0,55	0,66	0,66	0,49	0,46	0,15	0,86	0,78	0,82	0,78
	F-statistika	<b>0,73</b>	<b>0,65</b>	0,00	0,27	<b>0,72</b>	<b>0,65</b>	<b>0,62</b>	<b>0,72</b>	<b>0,70</b>	<b>0,56</b>	<b>0,53</b>	<b>0,30</b>	<b>0,89</b>	<b>0,82</b>	<b>0,85</b>	<b>0,82</b>
	PCA	0,71	0,61	0,00	0,22	0,69	0,61	0,58	0,70	0,68	0,52	0,49	0,22	0,88	0,80	0,84	0,80
1. odvod	brez	0,68	0,56	0,00	0,19	0,67	0,56	0,54	0,67	0,63	0,47	0,44	0,12	0,88	0,81	0,84	0,81
	varianca	0,66	0,55	0,00	0,19	0,64	0,56	0,53	0,64	0,62	0,47	0,43	0,11	0,86	0,80	0,82	0,80
	F-statistika	0,69	0,57	0,00	0,20	0,67	0,58	0,54	0,68	0,63	0,48	0,45	0,15	0,88	<b>0,82</b>	0,84	<b>0,82</b>
	PCA	0,68	0,56	0,00	0,19	0,67	0,56	0,54	0,67	0,63	0,48	0,44	0,13	0,88	0,81	0,84	0,81
SNV	brez	0,70	0,59	0,00	0,23	0,69	0,60	0,57	0,69	0,66	0,50	0,47	0,19	0,88	0,81	0,83	0,81
	varianca	0,65	0,54	0,00	0,18	0,64	0,54	0,51	0,64	0,62	0,45	0,42	0,08	0,86	0,78	0,81	0,78
	F-statistika	<b>0,73</b>	0,64	0,00	<b>0,28</b>	<b>0,72</b>	0,64	0,61	<b>0,72</b>	0,69	0,54	0,52	0,28	<b>0,89</b>	<b>0,82</b>	<b>0,85</b>	<b>0,82</b>
	PCA	0,69	0,58	0,00	0,22	0,68	0,59	0,56	0,68	0,65	0,49	0,46	0,17	0,87	0,80	0,83	0,80

Tabela B.4: **5-NN (10-CV)** Povprečne vrednosti mer za vrednotenje uvrščanja, pridobljene s 500 krat ponovljeno razdelitvijo na učno in testno množico pri uvrščanju z metodo 5-NN. Najboljši rezultat pri vsaki meri (najvišja vrednost v vsakem stolpcu) je zapisan krepko.



## Dodatek C

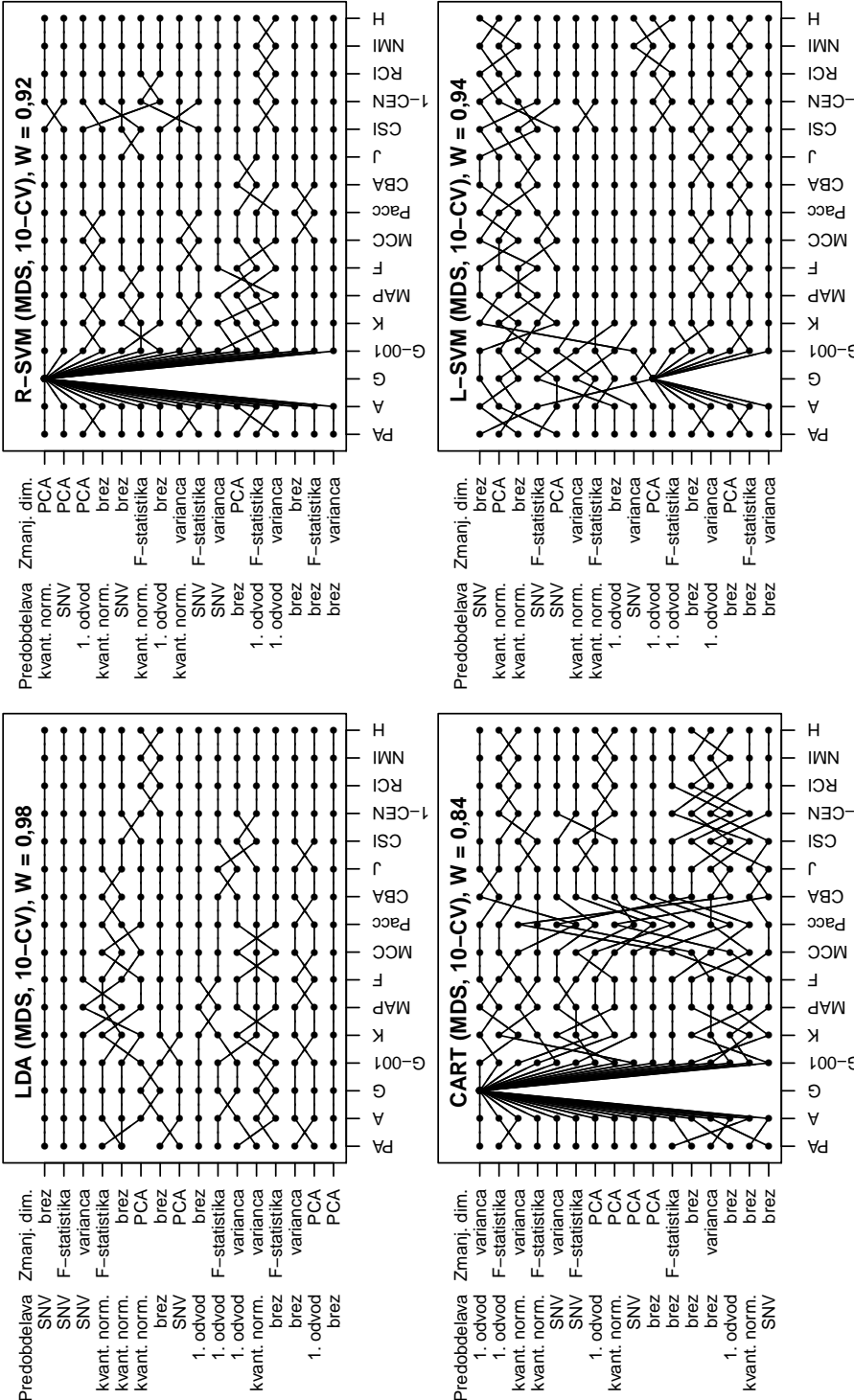
Primerjava mer vrednotenja  
uvrščanja za modele, ovrednotene  
s pristopom razdelitev na učno  
množico, ali z uporabo metode  
MDS



Slika C.1: Prikaz konkordance 16 mer za vrednotenje uvrščanja s Kendallovim koeficientom konkordance ( $W$ ) in s prikazom rangov na vzporednih oseh za modele uvrščanja, zgrajene z metodami LDA, CART, R-SVM in L-SVM ter ovrednotene s 500 krat ponovljeno razdelitvijo na učno in testno množico. Modeli so za vsako od metod uvrščanja razporejeni od najboljšega (zgoraj) do najslabšega (spodaj) glede na mero A.





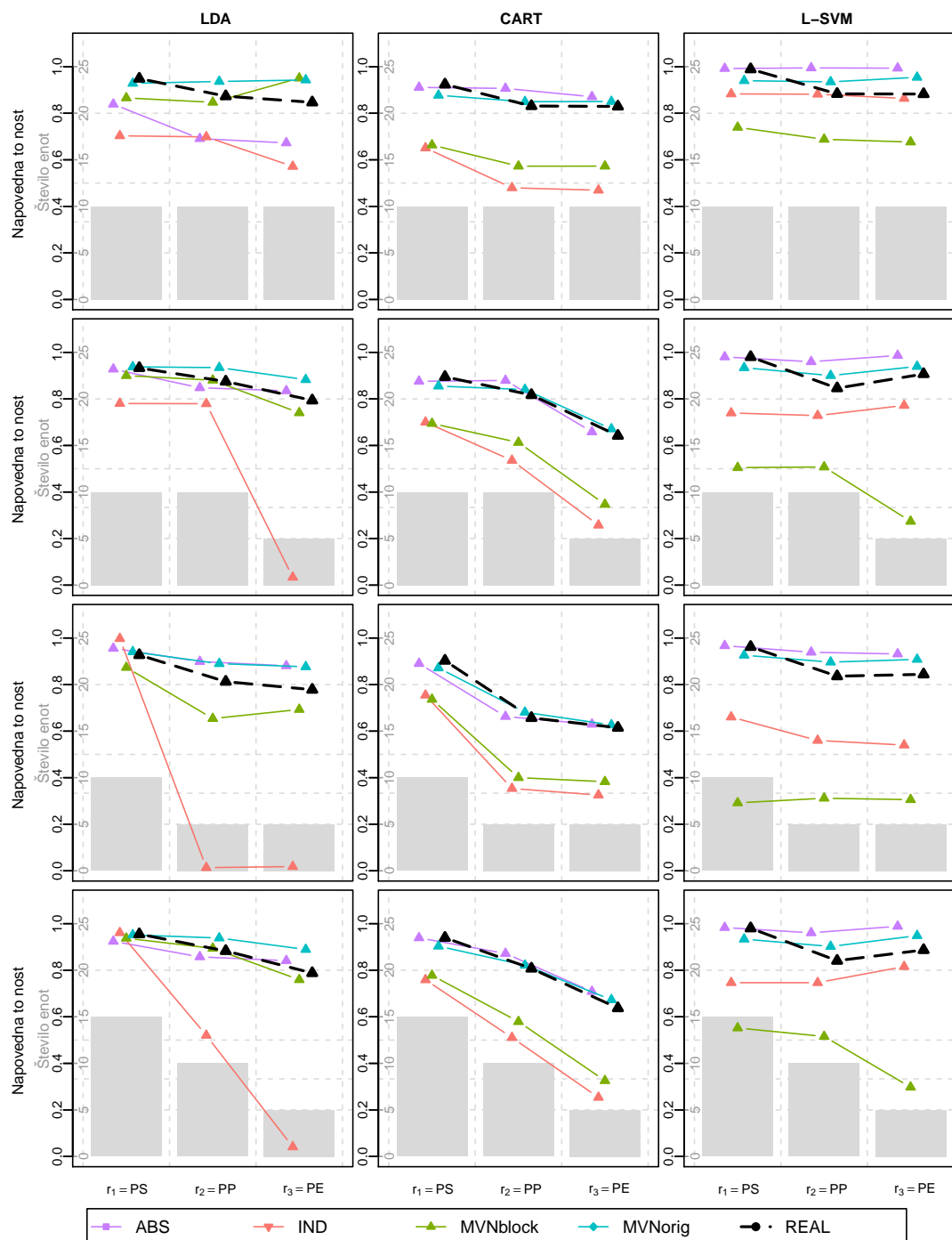


Slika C.3: Prikaz konkordance 16 mer za vrednotenje uvrščanja s Kendallovim koeficientom konkordance ( $W$ ) in s prikazom rangov na metodah LDA, CART, R-SVM in L-SVM. Pri katerih je bila za zmanjšanja neravnotežja uporabljen prečni preverjanjem s pregibanjem. Modeli so za vsako od metod uvrščanja razporejeni od najboljšega (zgoraj) do najslabšega (spodaj) glede na mero A.

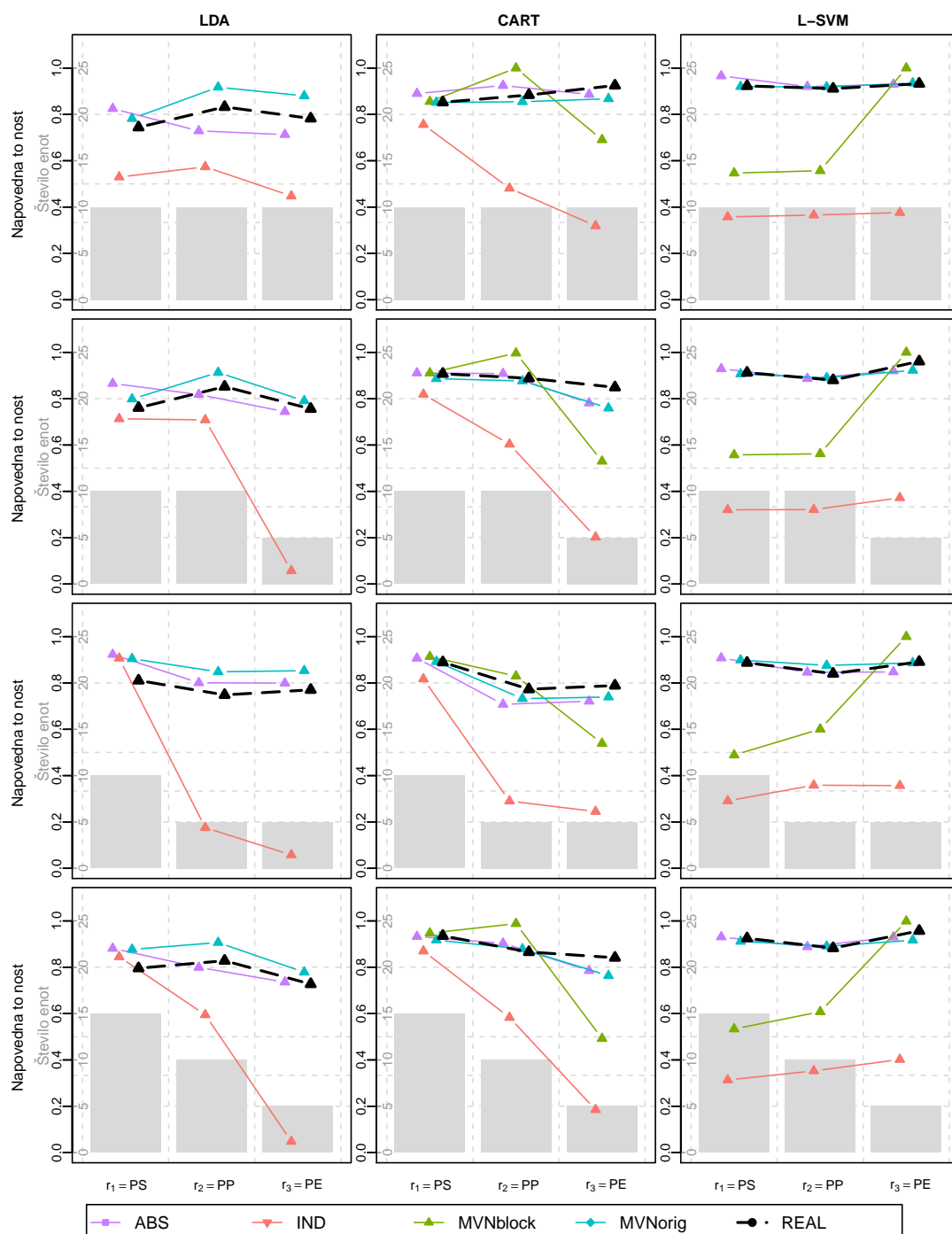
## Dodatek D

# Rezultati simulacij uvrščanja v tri razrede

### D.1 Predobdelave spektrov (Sliki D.1, D.2)

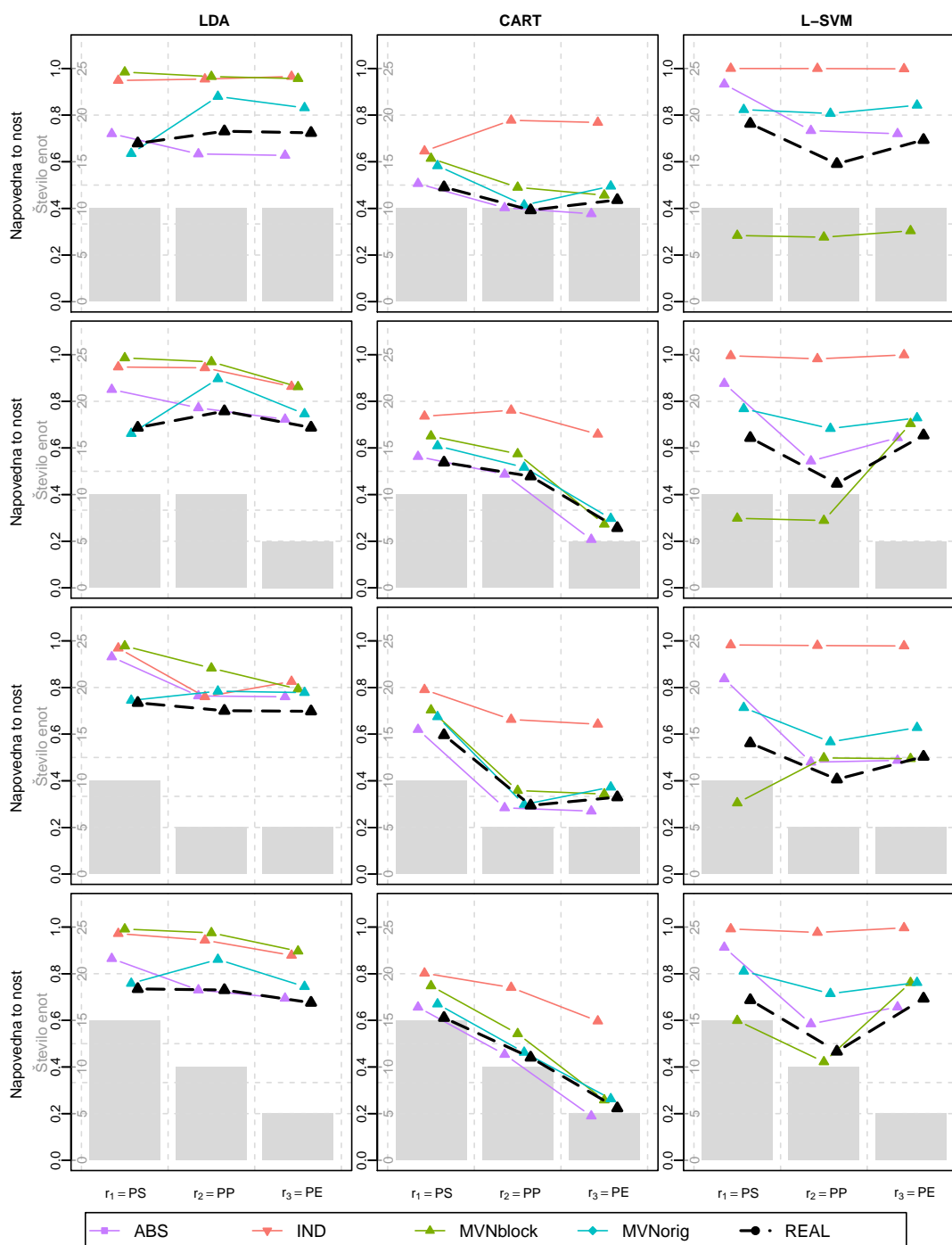


Slika D.1: **PS-PP-PE (SNV)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer so bile enote v prvem razredu iz skupine PS, v drugem razredu iz skupine PP in v tretjem razredu iz skupine PE. Razlike med skupinama PP in PE so bile manjše kot med PP in PS ali PE in PS. Enote so bile predobdelane z metodo SNV.

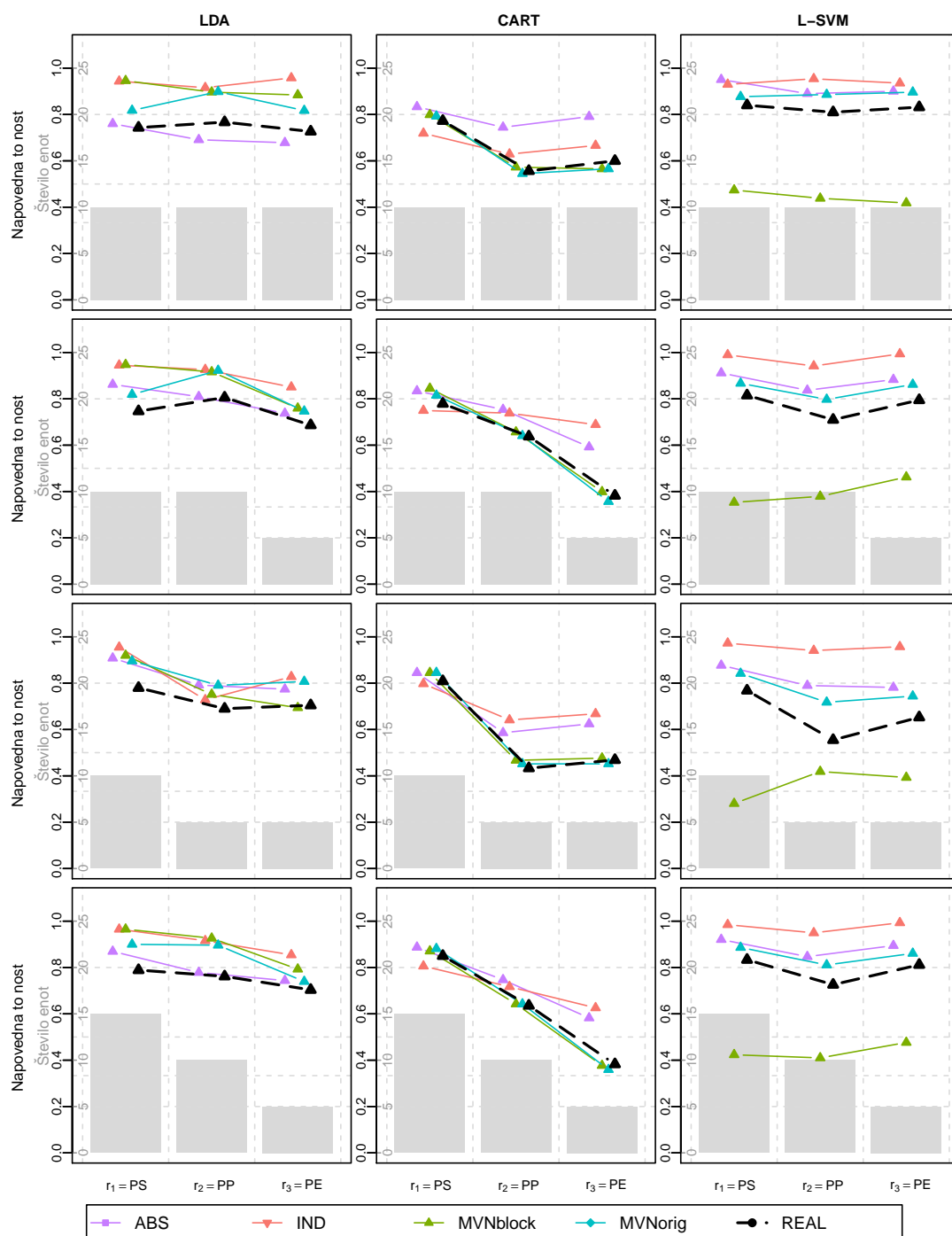


Slika D.2: **PS-PP-PE (1. odvod)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer so bile enote v prvem razredu iz skupine PS, v drugem razredu iz skupine PP in v tretjem razredu iz skupine PE. Razlike med skupinama PP in PE so bile manjše kot med PP in PS ali PE in PS. Na enotah je bil predhodno izračunan 1. odvod s pomočjo Savitzky-Golay filtra z oknom 7 in 1. stopnjo polinoma.

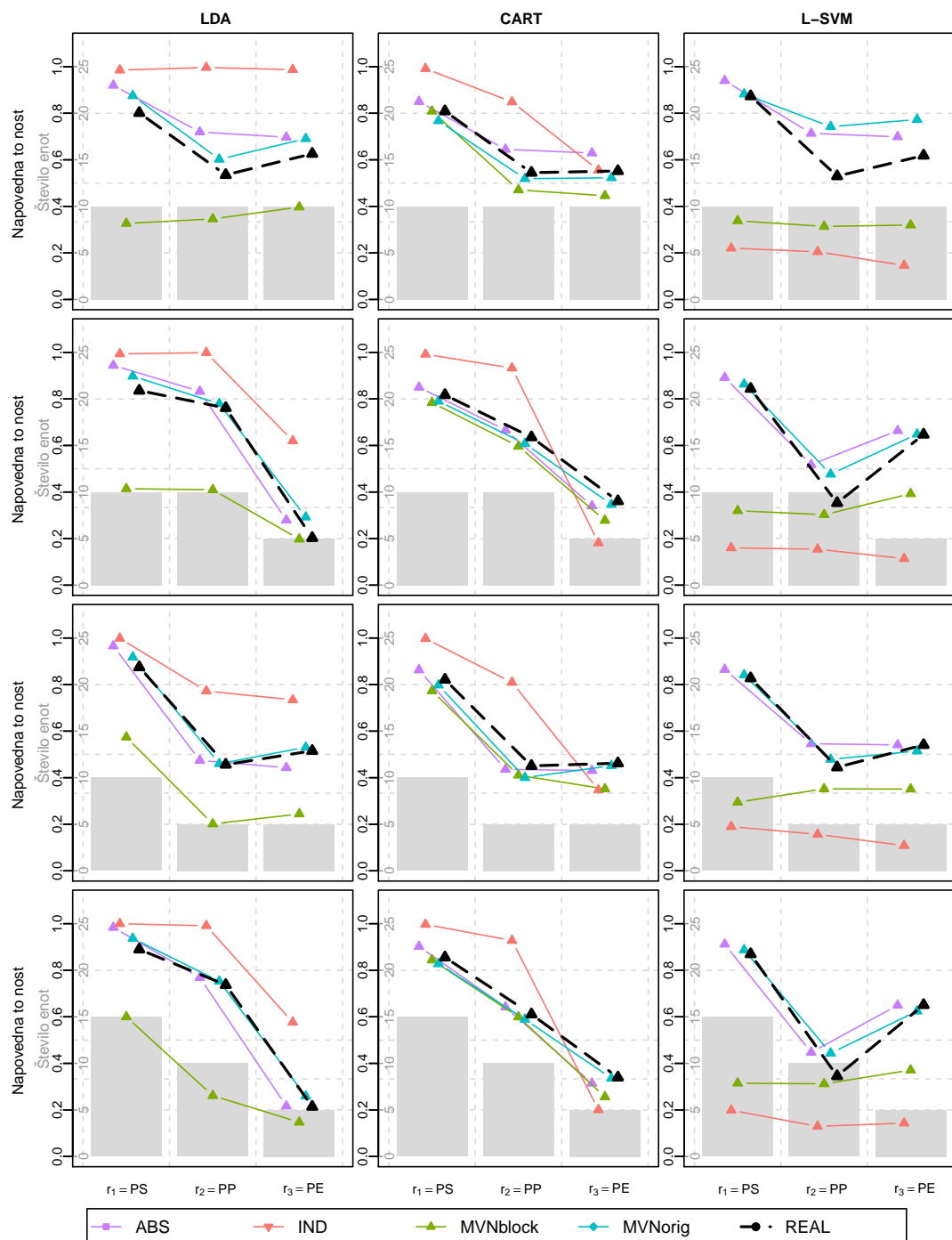
## D.2 Zmanjšanje dimenzije podatkov (Slike D.3, D.4, D.5)



Slika D.3: **PS-PP-PE (izbor glede na največjo varianco)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede (PS, PP, PE). Razlike med skupinama PP in PE so bile manjše kot med PP in PS ali PE in PS. Pri izgradnji klasifikatorja je bilo uporabljenih 50 spremenljivk z največjo varianco.



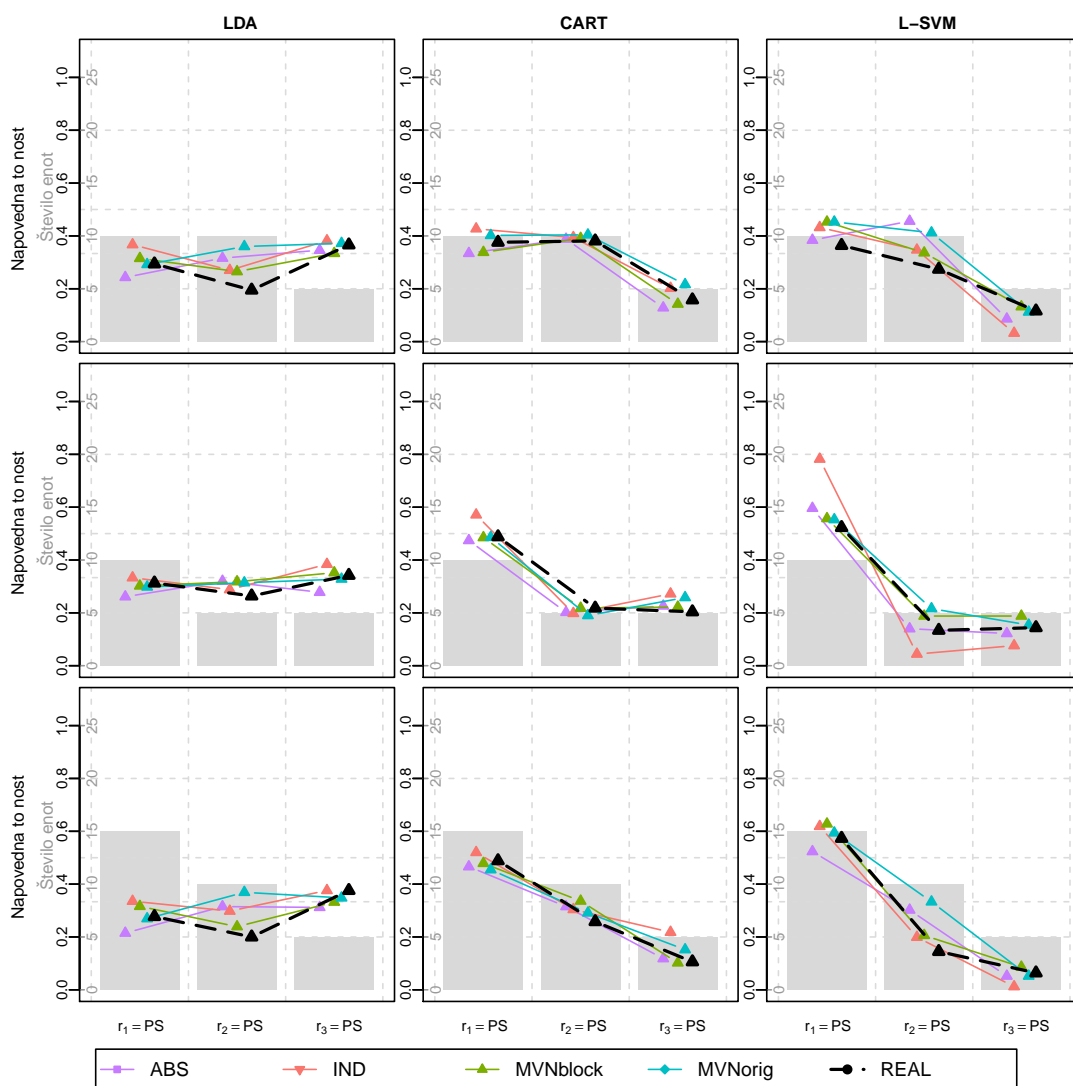
Slika D.4: **PS-PP-PE (izbor glede na največjo F-statistiko)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede (PS, PP, PE). Razlike med skupinama PP in PE so manjše kot med PP in PS ali PE in PS. Pri izgradnji klasifikatorja je bilo uporabljenih 50 spremenljivk z največjo F-statistiko.



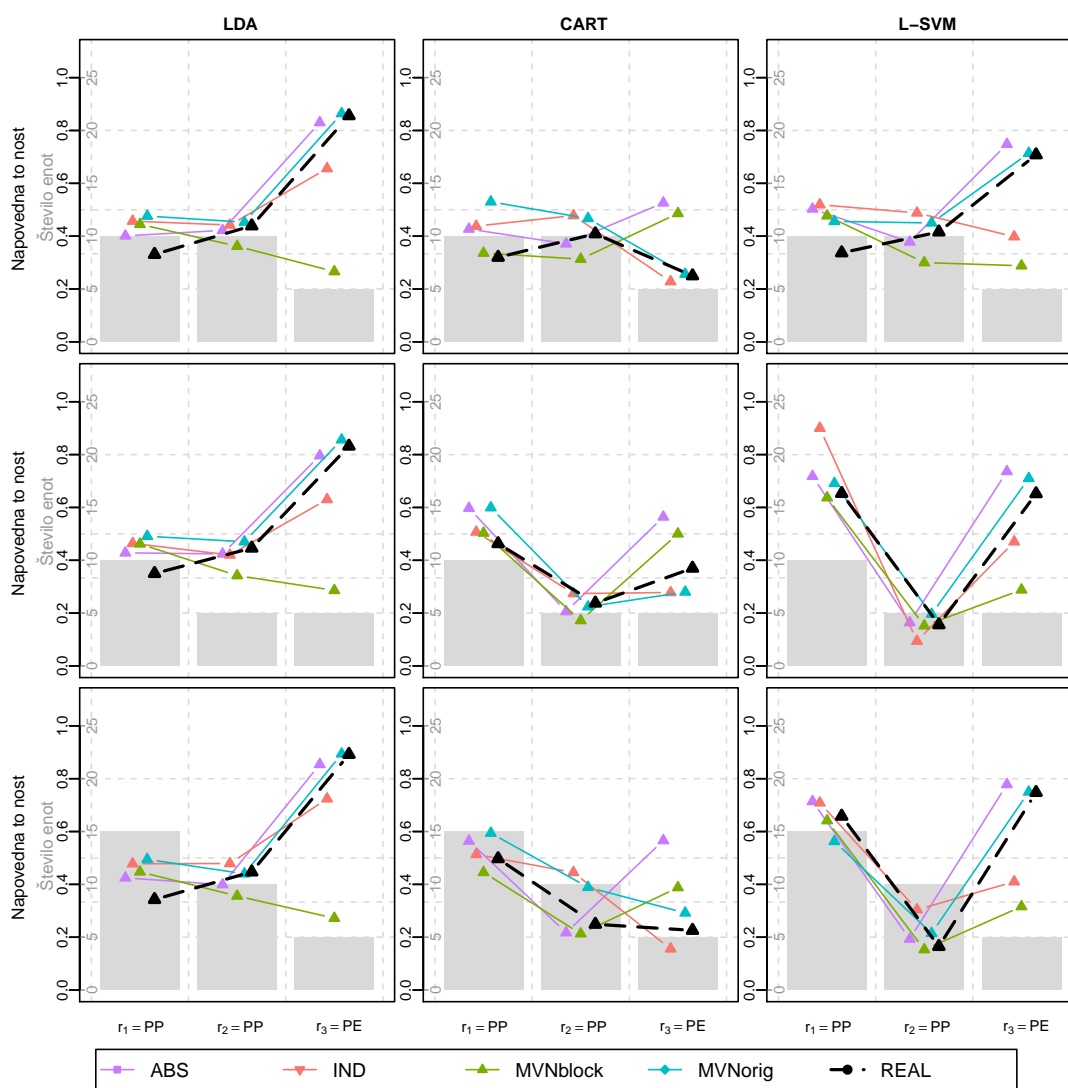
Slika D.5: **PS-PP-PE (PCA)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede (PS, PP, PE). Razlike med skupinama PP in PE so bile manjše kot med PP in PS ali PE in PS. Pri izgradnji klasifikatorja so bile uporabljene glavne komponente, izračunane po metodi PCA, ki so pokrile 99 % celotne variabilnosti.



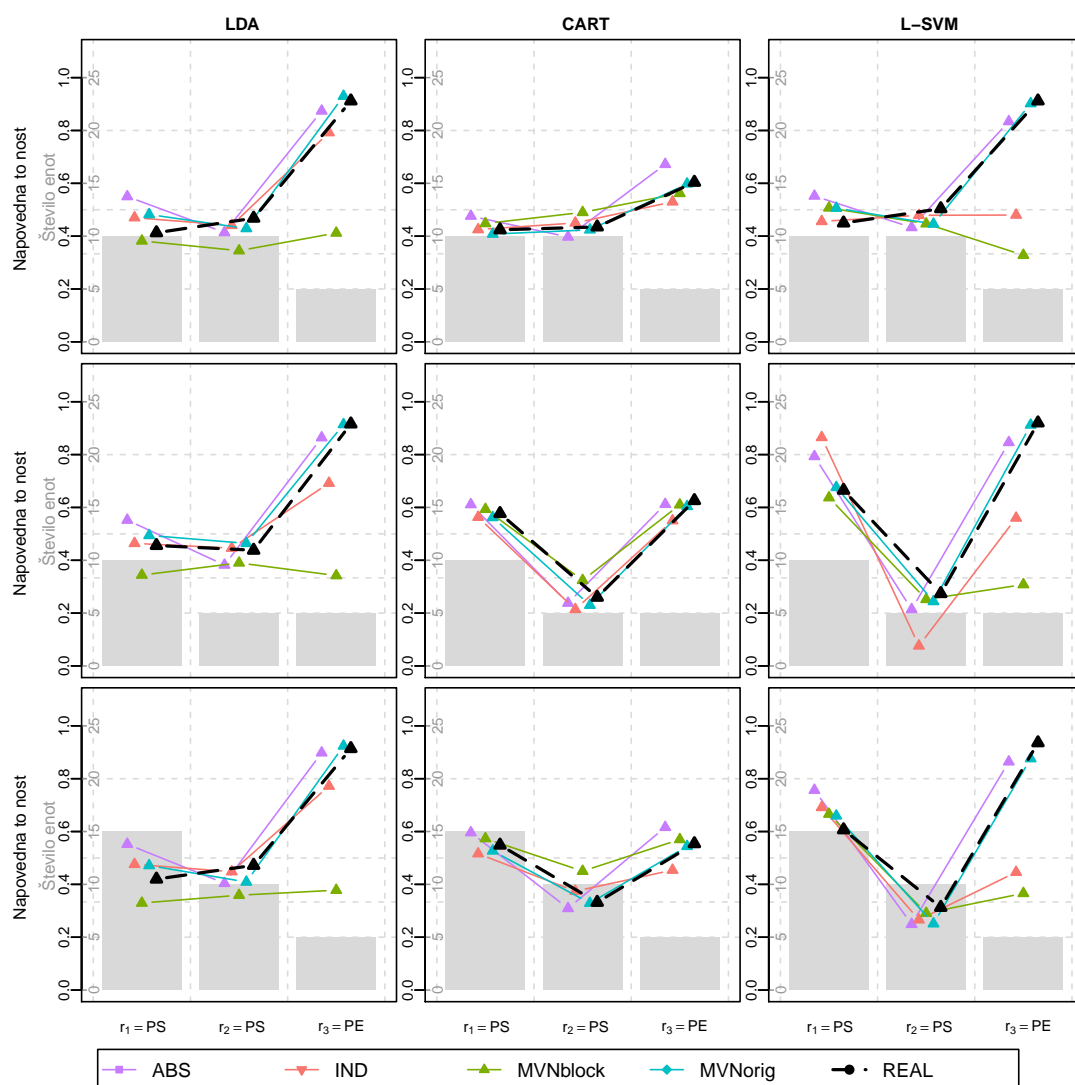
### D.3 Večkratno zmanjšanje večjega razreda (Slike D.6, D.7, D.8, D.9)



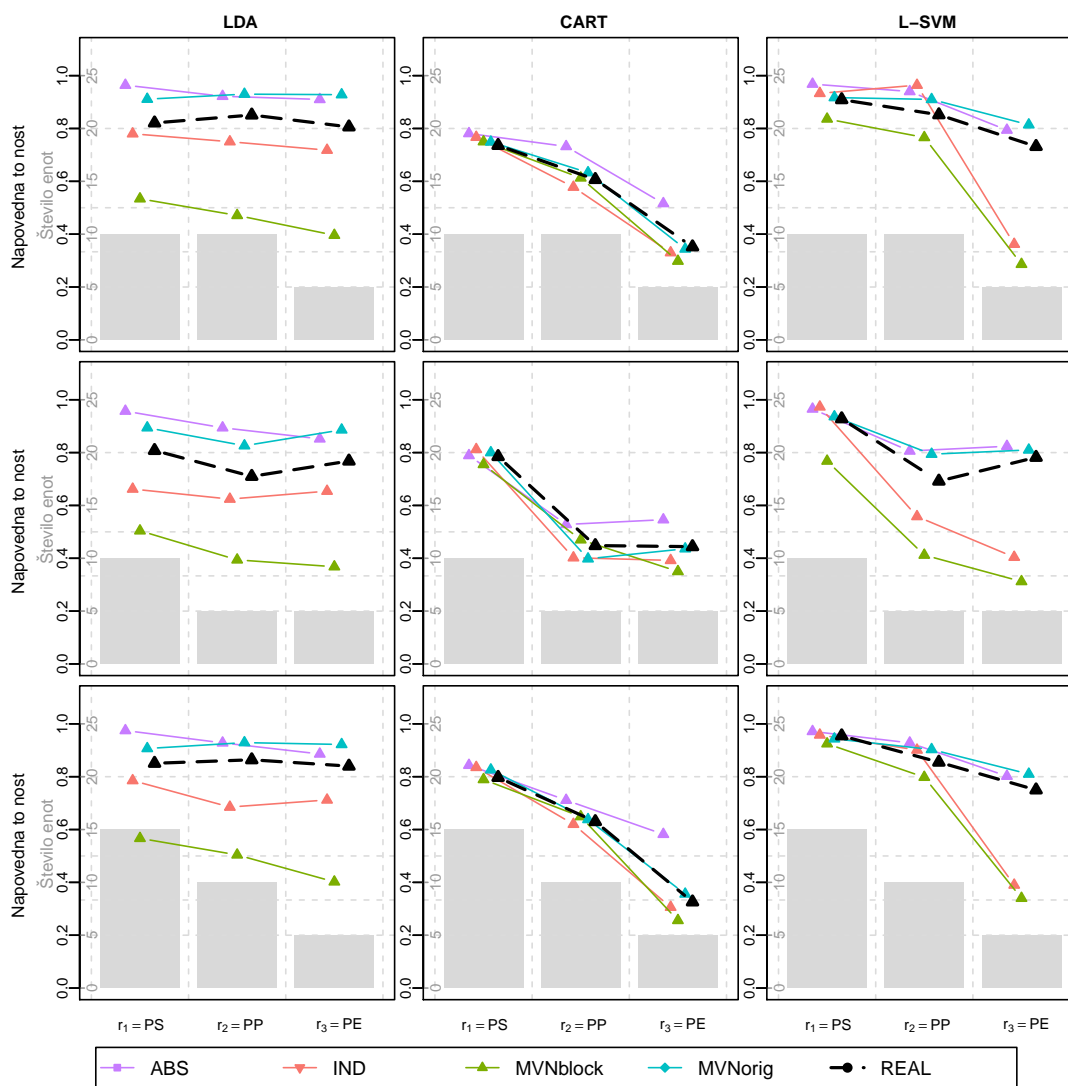
Slika D.6: **PS-PS-PS (MDS)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, med katerimi ni bilo razlik (PS, PS, PS). Pri gradnji klasifikatorja je bilo uporabljeno večkratno zmanjšanje večjega razreda v 100 ponovitvah.



Slika D.7: **PP-PP-PE (MDS)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer med prvima dvema razredoma ni bilo razlik (skupina PP), tretji razred se je z majhnimi razlikami razlikoval od prvih dveh (skupina PE). Pri gradnji klasifikatorja je bilo uporabljeno večkratno zmanjšanje večjega razreda v 100 ponovitvah.



Slika D.8: **PS-PS-PE (MDS)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede, pri čemer med prvima dvema razredoma ni razlik (skupina PS), tretji razred pa se je močno razlikoval od prvih dveh (skupina PE). Pri gradnji klasifikatorja je bilo uporabljeno večkratno zmanjšanje večjega razreda v 100 ponovitvah.



Slika D.9: **PS-PP-PE (MDS)**. Povprečne napovedne točnosti po razredih pri uvrščanju v tri razrede (PS, PP, PE). Razlike med skupinama PP in PE so manjše kot med PP in PS ali PE in PS. Pri gradnji klasifikatorja je bilo uporabljeno večkratno zmanjšanje večjega razreda v 100 ponovitvah.

## Dodatek E

# Pogosto uporabljene krajšave

Kratica	Slovenski izraz	Angleški izraz
A	A-povprečje	macro-average Aritmetic
ABS	podatki, generirani na podlagi teoretičnih absorpcij funkcionalnih skupin v NIR območju	
AUC	ploščina pod ROC krivuljo	Area Under the ROC Curve
BPNN	nevronska mreža z vzratnim širjenjem napake	Back Propagation Neural Network
CART	klasifikacijska in regresijska drevesa	Classification And Regression Trees
CBA	po razredih uravnotežena napovedna točnost	Class Balance Accuracy
CEN	razvrstitvena entropija	Confusion Entropy
CSI	indeks uspešnosti uvrščanja v posamezni razred	Individual Classification Success Index
CSI	indeks uspešnosti uvrščanja	Classification Success Index
CV	prečno preverjanje s pregibanjem	Cross-Validation
F	F-povprečje	mean F-measure
G	G-povprečje	macro-average Geometric, G-mean
H	normalizirana prenešana informacija	normalized transmitted information
IND	podatki, generirani neodvisno iz normalne porazdelitve	
J	Jaccardov koeficient	Jaccard's coefficient, mean intersection over union
k-NN	metoda k-najbližjih sosedov	<i>k</i> -Nearest Neighbours
LDA	linearna diskriminantna analiza	Linear Discriminant Analysis
LOOCV	prečno preverjanje z izpustitvijo ene enote	Leave One Out Cross-Validation
L-SVM	metoda podpornih vektorjev z linearno jedrno funkcijo	
MAP	makropovprečje napovedne vrednosti	Macro-averaged precision
MCC	korelacijski koeficient Matthew	Mathew's Correlation Coefficient, K-category correlation coefficient
MCS	multiplikativna korekcija spektra	Multiplicative Scatter Correction
MDS	večkratno zmanjšanje večjega razreda	Multiple Downsizing
MIR	srednje infrardeče območje	Mid Infrared
MVN	multivariatna normalna porazdelitev	Multivariate Normal distribution

MVNBloc	podatki, generirani iz multivariatne normalne porazdelitve z bločno kovariančno matriko	
MVNorig	podatki, generirani iz multivariatne normalne porazdelitve s parametri, ocenjenimi iz realnih podatkov	
NIR	bližnje infrardeče območje	Near Infrared
NIRS	bližnja infrardeča spektroskopija	Near Infrared Spectroscopy
NMI	normalizirana vzajemna informacija	Normalized Mutual Information
OAA	vsak proti vsem	One-Against-All
OAo	vsak proti vsakem	One-Against-One
PA	napovedna točnost	Predictive Accuracy
Pacc	verjetnostna napovedna točnost	Probabilistic accuracy measure
PCA	analiza glavnih komponent	Principal Component Analysis
PE	polietilen	Polyethylene
PLS	metoda delnih najmanjših kvadratov	Partial Least Squares
PLS-DA	diskriminantna analiza delnih najmanjših kvadratov	Partial Least Squares Discriminant Analysis
PNN	verjetnostna nevronska mreža	Probabilistic Neural Network
PP	polipropilen	Polypropylene
PS	polistiren	Polystyrene
PV	napovedna vrednost posameznega razreda	Predictive Value
RA	napovedna točnost naključnega klasifikatorja	Random Accuracy
RCI	relativna informacija klasifikatorja	Relative Classifier Information
ROC	krivulje ROC	Receiver Operating characteristic Curves
R-SVM	metoda podpornih vektorjev z radialno jedrno funkcijo	
RUS	zmanjšanje večjih razredov	Random Undersampling
SIMCA	mehko neodvisno modeliranje podobnosti po razredih	Soft Independent Modeling of Class Analogy
SMOTE		Synthetic Minority Over-sampling Technique
SNV	standardna normalna vektorska transformacija	Standard Normal Variate
S-PA	kombinacija občutljivosti in napovedne točnosti	Sensitivity-Accuracy approach
SVDD		Support Vector Data Description
SVM	metoda podpornih vektorjev	Support Vector Machines

Tabela E.1: Seznam pogosto uporabljenih kratic z angleškim in slovenskim pomenom, kjer obstajata.

# Literatura

- [1] D. A. Burns and E. W. Ciurczak, *Handbook of near-infrared analysis*. CRC press, 2007.
- [2] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, “A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies,” *Journal of Pharmaceutical and Biomedical Analysis*, vol. 44, pp. 683–700, July 2007.
- [3] D. Cozzolino, H. E. Smyth, and M. Gishen, “Feasibility Study on the Use of Visible and Near-Infrared Spectroscopy Together with Chemometrics To Discriminate between Commercial White Wines of Different Varietal Origins,” *J. Agric. Food Chem.*, vol. 51, pp. 7703–7708, Nov. 2003.
- [4] K. Thyholt and T. Isaksson, “Differentiation of Frozen and Unfrozen Beef Using Near-Infrared Spectroscopy,” *J. Sci. Food Agric.*, vol. 73, no. 4, pp. 525–532, 1997.
- [5] L. Esteve Agelet, D. D. Ellis, S. Duwick, A. S. Goggi, C. R. Hurburgh, and C. A. Gardner, “Feasibility of near infrared spectroscopy for analyzing corn kernel damage and viability of soybean and corn kernels,” *Journal of Cereal Science*, vol. 55, pp. 160–165, Mar. 2012.
- [6] V. R. Kondepati, H. M. Heise, and J. Backhaus, “Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy,” *Analytical and Bioanalytical Chemistry*, vol. 390, pp. 125–139, Jan. 2008.
- [7] M. Jamróiewicz, “Application of the near-infrared spectroscopy in the pharmaceutical technology,” *Journal of Pharmaceutical and Biomedical Analysis*, vol. 66, pp. 1–10, July 2012.
- [8] A. Sakudo, “Near-infrared spectroscopy for medical applications: Current status and future perspectives,” *Clinica Chimica Acta*, vol. 455, pp. 181–188, 2016.
- [9] R. M. Balabin and R. Z. Safieva, “Gasoline classification by source and type based on near infrared (NIR) spectroscopy data,” *Fuel*, vol. 87, pp. 1096–1101, June 2008.
- [10] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, “Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques,” *Analytica Chimica Acta*, vol. 671, pp. 27–35, June 2010.
- [11] W. van den Broek, E. Derks, E. van de Ven, D. Wienke, P. Geladi, and L. Buydens, “Plastic identification by remote sensing spectroscopic nir imaging using kernel partial least squares (kpls),” *Chemometrics and Intelligent Laboratory Systems*, vol. 35, no. 2, pp. 187–197, 1996.

- [12] W. van den Broek, D. Wienke, W. Melssen, and L. Buydens, "Plastic material identification with spectroscopic near infrared imaging and artificial neural networks," *Analytica Chimica Acta*, vol. 361, no. 1–2, pp. 161–176, 1998.
- [13] D. A. Cheng-Wen Chang, "Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties," *Soil Science Society of America Journal*, vol. 65, pp. 480–490, Mar. 2001.
- [14] E. D. Moreira, M. J. Pontes, R. K. Galvão, and M. C. Araújo, "Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection," *Talanta*, vol. 79, pp. 1260–1264, Oct. 2009.
- [15] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [16] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, and M. Hanpin, "Variables selection methods in near-infrared spectroscopy," *Analytica Chimica Acta*, vol. 667, no. 1–2, pp. 14–32, 2010.
- [17] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 27–46, 2013.
- [18] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–11, 2008.
- [19] L. Toloşi and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011.
- [20] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.
- [21] J. Weston and C. Watkins, "Multi-class support vector machines," tech. rep., Citeseer, 1998.
- [22] T. G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," Jan. 1995.
- [23] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.
- [24] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The Annals of Statistics*, vol. 26, pp. 451–471, Apr. 1998.
- [25] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," *Pattern Recognition*, vol. 40, pp. 4–18, Jan. 2007.
- [26] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, pp. 415–425, Mar. 2002.
- [27] H. He and E. A. Garcia, "Learning from Imbalanced Data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, pp. 1263–1284, Sept. 2009.



- [28] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, pp. 429–449, Oct. 2002.
- [29] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 11, no. 1, pp. 523+, 2010.
- [30] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Computational Intelligence and Data Mining, 2009. CIDM &#039;09. IEEE Symposium on*, pp. 324–331, IEEE, Mar. 2009.
- [31] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, pp. 20–29, June 2004.
- [32] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, pp. 63–77, Jan. 2006.
- [33] S. Wang and X. Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, pp. 1119–1130, Aug. 2012.
- [34] V. García, J. S. Sánchez, and R. A. Mollineda, "Classification of High Dimensional and Imbalanced Hyperspectral Imagery Data Pattern Recognition and Image Analysis," in *Pattern Recognition and Image Analysis* (J. Vitrià, J. a. M. Sanches, and M. Hernández, eds.), vol. 6669 of *Lecture Notes in Computer Science*, ch. 80, pp. 644–651, Berlin, Heidelberg: Springer Berlin / Heidelberg, 2011.
- [35] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, "Near-infrared (nir) spectroscopy for motor oil classification: From discriminant analysis to support vector machines," *Microchemical Journal*, vol. 98, no. 1, pp. 121–128, 2011.
- [36] O. Galtier, O. Abbas, Y. L. Dréau, C. Rebufa, J. Kister, J. Artaud, and N. Dupuy, "Comparison of pls1-da, pls2-da and {SIMCA} for classification by origin of crude petroleum oils by {MIR} and virgin olive oils by {NIR} for different spectral regions," *Vibrational Spectroscopy*, vol. 55, no. 1, pp. 132–140, 2011.
- [37] J. Zhao, H. Lin, Q. Chen, X. Huang, Z. Sun, and F. Zhou, "Identification of egg's freshness using {NIR} and support vector data description," *Journal of Food Engineering*, vol. 98, no. 4, pp. 408–414, 2010.
- [38] A. Candolfi, W. Wu, D. Massart, and S. Heuerding, "Comparison of classification approaches applied to nir-spectra of clinical study lots," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 16, no. 8, pp. 1329–1347, 1998.
- [39] M. Zoccola, N. Lu, R. Mossotti, R. Innocenti, and A. Montarsolo, "Identification of wool, cashmere, yak, and angora rabbit fibers and quantitative determination of wool and cashmere in blend: a near infrared spectroscopy study," *Fibers and Polymers*, vol. 14, no. 8, pp. 1283–1289, 2013.
- [40] M. Pontes, S. Santos, M. Araújo, L. Almeida, R. Lima, E. Gaião, and U. Souto, "Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry," *Food Research International*, vol. 39, no. 2, pp. 182–189, 2006.

- [41] M. Blanco, S. MasPOCH, I. Villarroja, X. Peralta, J. Gonzalez, and J. Torres, "Geographical origin classification of petroleum crudes from near-infrared spectra of bitumens," *Applied Spectroscopy*, vol. 55, no. 7, pp. 834–839, 2001.
- [42] L. Stothers, R. Guevara, and A. Macnab, "Classification of male lower urinary tract symptoms using mathematical modelling and a regression tree algorithm of noninvasive near-infrared spectroscopy parameters," *European Urology*, vol. 57, no. 2, pp. 327–333, 2010.
- [43] M. J. Cohen, J. P. Prenger, and W. F. DeBusk, "Visible-near infrared reflectance spectroscopy for rapid, nondestructive assessment of wetland soil quality," *Journal of Environmental Quality*, vol. 34, no. 4, pp. 1422–1434, 2005.
- [44] A. Fernández, V. López, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Systems*, vol. 42, pp. 97–110, Apr. 2013.
- [45] L. Lusa and R. Blagus, "The class-imbalance problem for high-dimensional class prediction," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 2, pp. 123–126, Dec 2012.
- [46] Q. Hai-bin, O. Dan-lin, and C. Yi-yu, "Background correction in near-infrared spectra of plant extracts by orthogonal signal correction," *Journal of Zhejiang University Science B*, vol. 6, no. 8, pp. 838–843, 2005.
- [47] Y. Ge, C. Morgan, J. Thomasson, and T. Waiser, "A new perspective to near-infrared reflectance spectroscopy: A wavelet approach," *Trans. ASABE*, vol. 50, no. 1, pp. 303–311, 2007.
- [48] G. Wang, M. Ma, Z. Zhang, Y. Xiang, and P. d. B. Harrington, "A novel dpso-svm system for variable interval selection of endometrial tissue sections by near infrared spectroscopy," *Talanta*, vol. 112, pp. 136–142, 2013.
- [49] B. Nadler and R. R. Coifman, "The prediction error in cls and pls: the importance of feature selection prior to multivariate calibration," *Journal of Chemometrics*, vol. 19, no. 2, pp. 107–118, 2005.
- [50] M. Shariati-Rad and M. Hasani, "Selection of individual variables versus intervals of variables in pls," *Journal of Chemometrics*, vol. 24, no. 2, pp. 45–56, 2010.
- [51] L. Xu, , and I. Schechter\*, "Wavelength selection for simultaneous spectroscopic analysis. experimental and theoretical study," *Analytical Chemistry*, vol. 68, no. 14, pp. 2392–2400, 1996.
- [52] B. Hemmateenejad and S. Karimi, "Construction of stable multivariate calibration models using unsupervised segmented principal component regression," *Journal of Chemometrics*, vol. 25, no. 4, pp. 139–150, 2011.
- [53] P. D. Wentzell and L. V. Montoto, "Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures," *Chemometrics and intelligent laboratory systems*, vol. 65, no. 2, pp. 257–279, 2003.
- [54] N. P. Bhatt, A. Mitna, and S. Narasimhan, "Multivariate calibration of non-replicated measurements for heteroscedastic errors," *Chemometrics and intelligent laboratory systems*, vol. 85, no. 1, pp. 70–81, 2007.

- [55] H.-W. Tan and S. D. Brown, "Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration," *Journal of Chemometrics*, vol. 16, no. 5, pp. 228–240, 2002.
- [56] M. Planck, "On the law of distribution of energy in the normal spectrum," *Annalen der Physik*, vol. 4, no. 553, p. 1, 1901.
- [57] J. Workman Jr and L. Weyer, *Practical guide and spectral atlas for interpretive near-infrared spectroscopy*. CRC Press, 2012.
- [58] P. M. Morse, "Diatomic molecules according to the wave mechanics. ii. vibrational levels," *Physical Review*, vol. 34, no. 1, p. 57, 1929.
- [59] G. Herzberg, *Molecular Spectra and Molecular Structure: II. Infrared and Raman Spectra of Polyatomic Molecules*. New York: Van Nostrand, 1945.
- [60] C. E. Miller, "Chemical principles of near-infrared technology," *Near-infrared technology in the agricultural and food industries*, vol. 2, 2001.
- [61] J. Coates, "Interpretation of infrared spectra, a practical approach," *Encyclopedia of analytical chemistry*, 2000.
- [62] G. Socrates, *Infrared and Raman characteristic group frequencies: tables and charts*. John Wiley & Sons, 2004.
- [63] P. Wu and H. Siesler, "The assignment of overtone and combination bands in the near infrared spectrum of polyamide 11," *Journal of Near Infrared Spectroscopy*, vol. 7, pp. 65–76, 1999.
- [64] M. Schwanninger, J. C. Rodrigues, and K. Fackler, "A review of band assignments in near infrared spectra of wood and wood components," *Journal of Near Infrared Spectroscopy*, vol. 19, no. 5, p. 287, 2011.
- [65] J. Workman and J. Workman, *Handbook of Organic Compounds: NIR, IR, Raman, and UV Spectra Featuring Polymers and Surfaces*, vol. 1. Academic, 2000.
- [66] Labsphere, *Reflectance materials and coatings*. Technical guide.
- [67] J. Yan, N. Villarreal, and B. Xu, "Characterization of degradation of cotton cellulosic fibers through near infrared spectroscopy," *Journal of Polymers and the Environment*, vol. 21, no. 4, pp. 902–909, 2013.
- [68] T. M. Abdul Rasheed, K. P. B. Moosad, V. P. N. Nampoori, and K. Sathianandan, "Overtone spectra of styrene and polystyrene in the visible and near infrared regions," *Pramana*, vol. 33, no. 3, pp. 391–395, 1989.
- [69] S. Lohumi, S. Lee, H. Lee, and B.-K. Cho, "A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration," *Trends in Food Science & Technology*, vol. 46, no. 1, pp. 85–98, 2015.
- [70] L.-G. Zhang, X. Zhang, L.-J. Ni, Z.-B. Xue, X. Gu, and S.-X. Huang, "Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy," *Food Chemistry*, vol. 145, pp. 342–348, 2014.

- [71] L. Chen, J. Wang, Z. Ye, J. Zhao, X. Xue, Y. V. Heyden, and Q. Sun, "Classification of chinese honeys according to their floral origin by near infrared spectroscopy," *Food Chemistry*, vol. 135, no. 2, pp. 338–342, 2012.
- [72] T. Woodcock, G. Downey, and C. P. O'Donnell, "Confirmation of Declared Provenance of European Extra Virgin Olive Oil Samples by NIR Spectroscopy," *J. Agric. Food Chem.*, vol. 56, pp. 11520–11525, Nov. 2008.
- [73] G. B. da Costa, D. D. S. Fernandes, A. A. Gomes, V. E. de Almeida, and G. Veras, "Using near infrared spectroscopy to classify soybean oil according to expiration date," *Food Chemistry*, vol. 196, pp. 539–543, 2016.
- [74] G. Vasques, J. Demattê, R. A. V. Rossel, L. Ramírez-López, and F. Terra, "Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths," *Geoderma*, vol. 223–225, pp. 73–78, 2014.
- [75] Y.-Z. Feng, G. Downey, D.-W. Sun, D. Walsh, and J.-L. Xu, "Towards improvement in classification of escherichia coli, listeria innocua and their strains in isolated systems based on chemometric analysis of visible and near-infrared spectroscopic data," *Journal of Food Engineering*, vol. 149, pp. 87–96, 2015.
- [76] N. M. Nawi, G. Chen, T. Jensen, and S. A. Mehdizadeh, "Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared," *Biosystems Engineering*, vol. 115, no. 2, pp. 154–161, 2013.
- [77] B. Chance, S. Nioka, J. Zhang, E. F. Conant, E. Hwang, S. Briest, S. G. Orel, M. D. Schnall, and B. J. Czerniecki, "Breast cancer detection based on incremental biochemical and physiological properties of breast cancers: A six-year, two-site study1," *Academic Radiology*, vol. 12, no. 8, pp. 925–933, 2005.
- [78] S. Wang, H. Chen, and X. Yao, "Negative correlation learning for classification ensembles," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, July 2010.
- [79] Åsmund Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.
- [80] M. J. Anzanello and F. S. Fogliatto, "A review of recent variable selection methods in industrial and chemometrics applications," *European Journal of Industrial Engineering*, vol. 8, no. 5, pp. 619–645, 2014.
- [81] W. Wu, B. Walczak, D. Massart, K. Prebble, and I. Last, "Spectral transformation and wavelength selection in near-infrared spectra classification," *Analytica Chimica Acta*, vol. 315, no. 3, pp. 243–255, 1995.
- [82] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, and M. Hanpin, "Variables selection methods in near-infrared spectroscopy," *Analytica Chimica Acta*, vol. 667, no. 1–2, pp. 14–32, 2010.
- [83] E. I. George, "The variable selection problem," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1304–1308, 2000.

- [84] B. Shi, B. Ji, D. Zhu, Z. Tu, and Z. Qing, "Study on genetic algorithms-based nir wavelength selection for determination of soluble solids content in fuji apples," *Journal of Food Quality*, vol. 31, no. 2, pp. 232–249, 2008.
- [85] F. Westad, A. Schmidt, and M. Kermit, "Incorporating chemical band-assignment in near infrared spectroscopy regression models," *Journal of Near Infrared Spectroscopy*, vol. 16, no. 3, pp. 265–273, 2008.
- [86] S. B. Kim, C. Temiyasathit, K. Bensalah, A. Tuncel, J. Cadeddu, W. Kabbani, A. V. Mathker, and H. Liu, "An effective classification procedure for diagnosis of prostate cancer in near infrared spectra," *Expert Systems with Applications*, vol. 37, no. 5, pp. 3863–3869, 2010.
- [87] R. Kohavi, *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. PhD thesis, Stanford University, Stanford, CA, USA, 1996. UMI Order No. GAX96-11989.
- [88] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2 of *IJCAI'95*, (San Francisco, CA, USA), pp. 1137–1143, Morgan Kaufmann Publishers Inc., 1995.
- [89] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings* (A. Sattar and B.-h. Kang, eds.), pp. 1015–1021, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- [90] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, pp. 412–424, May 2000.
- [91] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: An empirical analysis of supervised learning performance criteria," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04*, (New York, NY, USA), pp. 69–78, ACM, 2004.
- [92] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data – recommendations for the use of performance metrics," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 245–251, IEEE, 2013.
- [93] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "A survey on graphical methods for classification predictive performance evaluation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1601–1618, Nov 2011.
- [94] K. A. Spackman, "Signal detection theory: Valuable tools for evaluating inductive learning," in *Proceedings of the Sixth International Workshop on Machine Learning*, (San Francisco, CA, USA), pp. 160–163, Morgan Kaufmann Publishers Inc., 1989.
- [95] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

- [96] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, pp. 29–36, Apr. 1982.
- [97] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Chapman & Hall/CRC, 1 ed., Jan. 1984.
- [98] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [99] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.
- [100] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology (2nd edition)*. Sage Publications, 2004.
- [101] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*, vol. 999. MIT Press, 1999.
- [102] D. D. Lewis, "Evaluating text categorization," in *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, (Stroudsburg, PA, USA), pp. 312–318, Association for Computational Linguistics, 1991.
- [103] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.
- [104] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification.," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, 2009.
- [105] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, pp. 442–451, Oct. 1975.
- [106] J. Gorodkin, "Comparing two k-category assignments by a k-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, pp. 367–374, Dec. 2004.
- [107] M. Sigdel and R. S. Aygün, "Pacc - a discriminative and accuracy correlated measure for assessment of classification results," in *Machine Learning and Data Mining in Pattern Recognition: 9th International Conference, MLDM 2013, New York, NY, USA, July 19-25, 2013. Proceedings* (P. Perner, ed.), (Berlin, Heidelberg), pp. 281–295, Springer Berlin Heidelberg, 2013.
- [108] L. Mosley, *A balanced approach to the multi-class imbalance problem*. PhD thesis, Iowa State University, 2013.
- [109] L. Li, Y. Wu, and M. Ye, "Experimental comparisons of multi-class classifiers," *Informatica*, vol. 39, no. 1, 2015.
- [110] P. Jaccard, "The distribution of the flora in the alpine zone," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [111] S. Koukoulas and G. A. Blackburn, "Introducing new indices for accuracy evaluation of classified images representing semi-natural woodland environments," *Photogrammetric Engineering and Remote Sensing*, vol. 67, no. 4, pp. 499–510, 2001.

- [112] P. A. Gutiérrez, C. Hervás-Martínez, F. J. Martínez-Estudillo, and M. Carbonero, “A two-stage evolutionary algorithm based on sensitivity and accuracy for multi-class problems,” *Information Sciences*, vol. 197, pp. 20–37, 2012.
- [113] J.-M. Wei, X.-J. Yuan, Q.-H. Hu, and S.-Q. Wang, “A novel measure for evaluating classifiers,” *Expert Systems with Applications*, vol. 37, no. 5, pp. 3799–3809, 2010.
- [114] V. Sindhwani, P. Bhattacharya, and S. Rakshit, “Information theoretic feature crediting in multiclass support vector machines,” in *Proceedings of the 2001 SIAM International Conference on Data Mining*, pp. 1–18, SIAM, 2001.
- [115] A. D. Forbes, “Classification-algorithm evaluation: Five performance measures based on confusion matrices,” *Journal of Clinical Monitoring*, vol. 11, no. 3, pp. 189–206, 1995.
- [116] T. D. Wickens, *Multiway contingency tables analysis for the social sciences*. Psychology Press, 2014.
- [117] N. Abramson, *Information theory and coding*. New York: McGraw-Hill, 1963.
- [118] T. Kreuz, J. S. Haas, A. Morelli, H. D. Abarbanel, and A. Politi, “Measuring spike train synchrony,” *Journal of Neuroscience Methods*, vol. 165, no. 1, pp. 151–161, 2007.
- [119] R. M. Fano and D. Hawkins, “Transmission of information: A statistical theory of communications,” *American Journal of Physics*, vol. 29, no. 11, pp. 793–794, 1961.
- [120] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
- [121] International Conference on Cognitive Science, *Recall & Precision versus The Book-maker*, (Sydney, Australia), 2003.
- [122] T. Fawcett, “Using rule sets to maximize roc performance,” in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 131–138, 2001.
- [123] D. J. Hand and R. J. Till, “A simple generalisation of the area under the roc curve for multiple class classification problems,” *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.
- [124] S. Wu, P. Flach, and C. Ferri, “An improved model selection heuristic for auc,” in *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings* (J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, eds.), pp. 478–489, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
- [125] C. T. Nakas and C. T. Yiannoutsos, “Ordered multiple-class roc analysis with continuous measurements,” *Statistics in Medicine*, vol. 23, no. 22, pp. 3437–3449, 2004.
- [126] G. Lebanon and J. D. Lafferty, “Cranking: Combining rankings using conditional probability models on permutations,” in *Proceedings of the Nineteenth International Conference on Machine Learning, ICML ’02*, (San Francisco, CA, USA), pp. 363–370, Morgan Kaufmann Publishers Inc., 2002.
- [127] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

- [128] I. J. Good, “Rational decisions,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 14, no. 1, pp. 107–114, 1952.
- [129] I. J. Good, “Corroboration, explanation, evolving probability, simplicity and a sharpened razor,” *The British Journal for the Philosophy of Science*, vol. 19, no. 2, pp. 123–143, 1968.
- [130] P. Flach and E. T. Matsubara, “A simple lexicographic ranker and probability estimator,” in *Proceedings of the 18th European Conference on Machine Learning, ECML ’07*, (Berlin, Heidelberg), pp. 575–582, Springer-Verlag, 2007.
- [131] R. Caruana and A. Niculescu-Mizil, “Data mining in metric space: An empirical analysis of supervised learning performance criteria,” in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’04*, (New York, NY, USA), pp. 69–78, ACM, 2004.
- [132] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, “Predicting the clinical status of human breast cancer by using gene expression profiles,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11462–11467, 2001.
- [133] G. A. Miller and P. E. Nicely, “An analysis of perceptual confusions among some english consonants,” *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.
- [134] M. Meilă, “Comparing clusterings—an information based distance,” *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.
- [135] F. J. Valverde-Albacete and C. Peláez-Moreno, “Two information-theoretic tools to assess the performance of multi-class classifiers,” *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1665–1671, 2010. Pattern Recognition of Non-Speech Audio.
- [136] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [137] R. Alejo, J. A. Antonio, R. M. Valdovinos, and J. H. Pacheco-Sánchez, “Assessments metrics for multi-class imbalance learning: A preliminary study,” in *Pattern Recognition: 5th Mexican Conference, MCPR 2013, Querétaro, Mexico, June 26-29, 2013. Proceedings* (J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. S. Rodríguez, and G. S. di Baja, eds.), (Berlin, Heidelberg), pp. 335–343, Springer Berlin Heidelberg, 2013.
- [138] G. Jurman, S. Riccadonna, and C. Furlanello, “A comparison of mcc and cen error measures in multi-class prediction,” *PloS one*, vol. 7, no. 8, p. e41882, 2012.
- [139] “Preservation of plastic artefacts in museum collections. european community’s seventh framework programme, grant agreement no. 212218.” <http://popart-highlights.mnhn.fr/index.html.htm>. Accessed: 2015-10-07.
- [140] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [141] B. M. Bolstad, *preprocessCore: A collection of pre-processing functions*, 2014. R package version 1.28.0.



- [142] Signal developers, *signal: Signal processing*, 2013.
- [143] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [144] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, 4 ed., 2002. ISBN 0-387-95457-0.
- [145] A. Beygelzimer, S. Kakadet, J. Langford, S. Arya, D. Mount, and S. Li, *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2013. R package version 1.1.
- [146] T. Therneau, B. Atkinson, and B. Ripley, *rpart: Recursive Partitioning and Regression Trees*, 2015. R package version 4.1-9.
- [147] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien, 2014. R package version 1.6-4.
- [148] R. Barnes, M. Dhanoa, and S. J. Lister, “Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra,” *Applied spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.
- [149] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,” *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.
- [150] A. Savitzky and M. J. Golay, “Smoothing and differentiation of data by simplified least squares procedures,” *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [151] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.
- [152] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [153] E. Fix and J. L. Hodges Jr, “Discriminatory analysis-nonparametric discrimination: consistency properties,” tech. rep., DTIC Document, 1951.
- [154] E. Fix and J. L. Hodges, “Discriminatory analysis. nonparametric discrimination: consistency properties,” *International Statistical Review/Revue Internationale de Statistique*, pp. 238–247, 1989.
- [155] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [156] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [157] D. Tao and X. Tang, “Random sampling based svm for relevance feedback image retrieval,” in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. 647–652, June 2004.
- [158] C. E. Shannon, “A Mathematical Theory of Communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

- [159] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiology*, vol. 31, no. 4, pp. 306–315, 2007.
- [160] R. Baeza-Yates and B. Ribeiro-Neto, "Modern information retrieval addison-wesley longman," *Reading MA*, 1999.
- [161] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503–514, 1989.
- [162] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.
- [163] S. Siegel and N. Castellan, *Nonparametric statistics for the behavioral sciences*. McGraw–Hill, Inc., 2 ed., 1988.
- [164] D. Wilkie, "Pictorial representation of kendalls rank correlation coefficient," *Teaching Statistics*, vol. 2, no. 3, pp. 76–78, 1980.
- [165] G. Vidmar, *Prikaz večrazsežnih podatkov s konveksnolupinskimi in konkordančnimi diagrami*. PhD thesis, Medicinska fakulteta, Ljubljana, 3 2007.
- [166] W.-J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Briefings in Bioinformatics*, 2012.
- [167] V. Šuštar, J. Kolar, L. Lusa, T. Learner, M. Schilling, R. Rivenc, H. Khanjian, and D. Koleša, "Identification of historical polymers using near-infrared spectroscopy," *Polymer Degradation and Stability*, vol. 107, pp. 341–347, 2014.
- [168] T. Mehta, M. Tanik, and D. B. Allison, "Towards sound epistemological foundations of statistical methods for high-dimensional biology," *Nature genetics*, vol. 36, no. 9, pp. 943–947, 2004.
- [169] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature reviews genetics*, vol. 7, no. 1, pp. 55–65, 2006.
- [170] G. L. Gadbury, Q. Xiang, L. Yang, S. Barnes, G. P. Page, and D. B. Allison, "Evaluating statistical methods using plasmode data sets in the age of massive public databases: An illustration using false discovery rates," *PLOS Genetics*, vol. 4, pp. 1–8, 6 2008.
- [171] D. C. Paranagama, *Correlation and variance stabilization in the two group comparison case in high dimensional data under dependencies*. PhD thesis, Kansas State University, 2011.
- [172] J. M. Franklin, S. Schneeweiss, J. M. Polinski, and J. A. Rassen, "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases," *Computational Statistics & Data Analysis*, vol. 72, pp. 219–226, 2014.
- [173] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

- 
- [174] T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, pp. 552–568, May 2011.
- [175] M. Y. Park, T. Hastie, and R. Tibshirani, "Averaged gene expressions for regression," *Biostatistics*, vol. 8, no. 2, p. 212, 2006.
- [176] M. Sigdel and R. Aygün, "Pacc - A Discriminative and Accuracy Correlated Measure for Assessment of Classification Results," in *Machine Learning and Data Mining in Pattern Recognition* (P. Perner, ed.), vol. 7988 of *Lecture Notes in Computer Science*, pp. 281–295, Springer Berlin Heidelberg, 2013.