Univerza v Mariboru

Fakulteta za elektrotehniko, računalništvo in informatiko

DOKTORSKA DISERTACIJA

NADOMEŠČANJE MANJKAJOČIH VREDNOSTI S POMOČJO ROTACIJSKEGA REGRESIJSKEGA GOZDA

Miroslav Palfy

Mentor: prof. dr. Peter Kokol

Somentor: izred. prof. dr. Milan Zorman

Maribor, december 2009

UDK: 004.89:004.9(043.3)

Z iskreno zahvalo

mentorju prof. dr. Petru Kokolu, ki me je vrsto let vodil na poti podiplomskega študija ter s svojim optimizmom in spodbudo bistveno pripomogel k nastanku tega dela,

somentorju izred. prof. dr. Milanu Zormanu, s čigar pomočjo sem uspešno rešil marsikatero težavo,

prof. dr. Zmagu Turku, dr. med., da sem se lahko kot mladi raziskovalec pod njegovim mentorstvom posvetil študijskemu delu in se vedno obrnil nanj po nasvet

ter dr. Gregorju Štiglicu, ki mi je predstavil rotacijski gozd in mi nesebično pomagal izoblikovati temo doktorske disertacije.

Posebno hvaležen sem ženi Marjani, ki mi je ves čas stala ob strani, me priganjala k delu, tudi sama pomagala, ko sem se znašel v časovni stiski in vedno imela neomejeno mero razumevanja.

Hvala tudi

vsem drugim, ki so s svojo pomočjo kakorkoli sodelovali pri ustvarjanju te naloge, sodelavcem in kolegom, predvsem pa staršem in bratu, ki so me vedno podpirali pri mojem študiju.

KAZALO

Nadomeščanje manjkajočih vrednosti s pomočjo rotacijskega regresijskega gozda

UDK: 004.89:004.9(043.3)

Ključne besede: strojno učenje, rotacijski gozd, nadomeščanje manjkajočih vrednosti, regresijsko drevo, ansambel regresorjev

Povzetek:

Manjkajoče vrednosti predstavljajo pogosto težavo, ki spremlja ustvarjanje podatkovnih baz, bodisi če se podatki zbirajo s pomočjo anket bodisi če so pridobljeni iz načrtovanih eksperimentov. Ne glede na to, koliko truda je vloženo za zagotavljanje popolne izpolnjenosti vprašalnikov ali v skrbno načrtovanje znanstvenega poskusa, se manjkajočim vrednostim pogosto ni možno izogniti. Nepopolni podatki so, odvisno od razmerja v katerem se pojavljajo manjkajoče vrednosti, lahko neustrezni za nadaljnjo analizo, medtem ko je brisanje vzorcev z manjkajočimi vrednostmi, posebno ko njihov odstotek ni dovolj majhen in ti vzorci predstavljajo pomembne informacije, lahko zelo neustrezno. Za reševanje tega problema se tako na področju statistične analize uporabljajo različne metode za nadomeščanje manjkajočih vrednosti.

Z namenom zapolnitve vrzeli, ki obstaja med obstoječimi metodami enkratnega vstavljanja manjkajočih vrednosti in modeli, ki temeljijo na večkratnem vstavljanju in pri katerih je za vsak cikel vstavljanja potrebna ločena statistična analiza, smo v okviru disertacije razvili nov postopek nadomeščanja manjkajočih vrednosti, ki temelji na ansambelskem pristopu nadzorovanega strojnega učenja. Uporabili smo ansambel, imenovan rotacijski regresijski gozd, ki predstavlja varianto rotacijskega gozda (Rotation forest), kot so ga razvili Rodríguez, Kuncheva in Alonso (Rodríguez, Kuncheva, & Alonso, 2006), pri katerem smo namesto osnovne metode, namenjene reševanju klasifikacijskih problemov, uporabili modelno regresijsko drevo.

Našo metodo za nadomeščanje manjkajočih vrednosti smo primerjali z 9 drugimi popularnimi metodami, pri čemer smo merili natančnost metod in njihovo sposobnost ohranjanja variance po vstavljanju različnih deležev manjkajočih vrednosti. Meritve smo izvedli na 14 javno dostopnih podatkovnih množicah in eni umetno ustvarjeni množici, tako da smo obravnavali vse mehanizme nastanka manjkajočih vrednosti, kot jih je definiral Rubin (Rubin, 1976).

Na podlagi poizkusov smo ugotovili, da naša metoda v povprečju natančneje napoveduje manjkajoče vrednosti v izbranih podatkovnih množicah, ne glede na mehanizem nastanka manjkajočih vrednosti. Prav tako smo pokazali, da z uvedbo dodatne stohastične metode za ohranjanje variance naš rotacijski regresijski gozd bolje ohranja varianco od vseh preostalih metod, ki izvajajo enkratno vstavljanje, pri čemer po svoji natančnosti še vedno prekaša vse metode.

V disertaciji smo v uvodnih, teoretičnih poglavjih podrobneje opisali problematiko manjkajočih vrednosti ter obstoječe metode, ki se najpogosteje uporabljajo za njihovo nadomeščanje. Predstavili smo rotacijski regresijski gozd in stohastično metodo za ohranjanje variance. Največjo pozornost smo posvetili rezultatom poizkusov, na podlagi katerih smo v zaključku izoblikovali priporočila za uporabo rotacijskega regresijskega gozda za nadomeščanje manjkajočih vrednosti ter predstavili izhodišča za nadaljnje delo.

Missing values imputation using a rotation regression forest

UDC: 004.89:004.9(043.3)

Keywords: machine learning, rotation forest, missing value imputation, regression tree, ensemble of regressors

Abstract:

Missing values represent a common problem, plaguing many databases; either based on surveys and questionnaires or designed experiments. No matter how carefully the surveys are taken, or how well the experiments are designed, missing values can occur. Incomplete data can, depending on the amount of missing values, be unsuitable for further statistical analysis, while case deletion, especially when dealing with considerable amounts of missing values, can be very inappropriate. Therefore different methods were developed which can be used to impute missing data.

The main goal of this dissertation was to develop a new imputation method, which would narrow the gap between single-impute methods and multiple-imputation models, which require standard statistical analysis to be carried out on multiple imputed data sets. For this purpose we used an ensemble-based approach to supervised machine learning. We relied on a variation of rotation forest ensemble, developed by Rodríguez, Kuncheva and Alonso (Rodríguez, Kuncheva, & Alonso, 2006) which we named “rotation regression forest”, since we used a model regression tree as a base method instead of a method used for classification purposes.

We selected 9 other popular imputation methods for comparison with our ensemble where we measured their accuracy as well as their ability to preserve the variance structure within data when dealing with different amounts of missing values. Measurements were carried out on 14 different public access datasets and one artificial dataset to account for each of the three missingness mechanisms, as described by Rubin (Rubin, 1976).

Based on results of these tests we concluded that, on average, our method is the most accurate among the selected methods, no matter which misingness mechanism is responsible for missing values. When an additional stochastic method for preservation of variance was used, our rotation regression forest was able to preserve the variance structure within data better than any other single-impute method, while still besting them all in accuracy.

The introductory, more theoretical chapters of this dissertation deal with supervised machine learning, missing values and commonly used imputation methods. Rotation regression forest ensemble was introduced, as well as our stochastic method for preservation of variance. The bulk of our work is focused on results, gained through empirical experiments, which were used to model our recommendations concerning the use of rotation regression forest ensemble for imputation of missing values and to form starting points for possible future work.

1 Uvod

1.1 Motivacija

Pogosto težavo, ki spremlja ustvarjanje podatkovnih baz, bodisi če se podatki zbirajo s pomočjo anket (Acock, 2005) bodisi če so pridobljeni iz načrtovanih eksperimentov (Bennett, 2007), predstavljajo manjkajoče vrednosti. Ne glede na to, koliko truda je vloženo za zagotavljanje popolne izpolnjenosti vprašalnikov ali v skrbno načrtovanje znanstvenega poskusa, se manjkajočim vrednostim pogosto ni možno izogniti. Konkretni odgovori ali meritve lahko manjkajo, ker oseba ni želela odgovoriti na določeno vprašanje oziroma meritev ni bila izvedljiva, ali pa so pridobljene vrednosti v nekaterih primerih nesmiselne ali nenamerno izpuščene (šum) (Allison, 2001). Kot primer omenimo le področje bioinformatike, kjer smo danes priča izjemno obsežnim podatkovnim bazam, pridobljenim iz mikromrež (micorarrays) (Troyanskaya, in drugi, 2001). Še do nedavna izjemno dragi in kompleksni postopki meritve ekspresije genov mnogokrat niso ponovljivi iz ekonomskih razlogov ali zaradi nedostopnosti biološkega vzorca (Bo, Dysvik, & Jonassen, 2004), medtem ko so se manjkajoče vrednosti lahko pojavile kot posledica praha ali prask na steklu oziroma napak med tiskanjem ali hibridizacijo (Sehgal, Gondal, & Dooley, 2005). Nepopolni podatki so, odvisno od razmerja v katerem se pojavljajo manjkajoče vrednosti, lahko neustrezni za nadaljnjo analizo. Little in Rubin sta, med drugimi, prikazala, kako neustrezno je lahko brisanje vzorcev z manjkajočimi vrednostmi, posebno ko njihov odstotek ni dovolj majhen in ti vzorci predstavljajo pomembne informacije (Little & Rubin, 1987). V takšnih okoliščinah lahko brisanje vzorcev privede do občutne pristranskosti rezultatov. Prav tako lahko zmanjšanje velikosti baze zmanjša tudi statistično moč analize (Schafer, 1997). Za reševanje tega problema se tako na področju statistične analize uporabljajo različne metode (Horton & Kleinman, 2007).

Tradicionalni pristopi so relativno preprosti in vključujejo brisanje vzorcev ali vstavljanje povprečnih vrednosti. Slednje ima lahko še posebej negativen učinek, saj lahko močno vpliva na variabilnost podatkov (Scheffer, 2002). V zadnjih dveh desetletjih je bila pozornost usmerjena predvsem v regresijske modele (Segal, 1988) in vstavljanje vrednosti s pomočjo algoritma EM (Expectation - Maximisation algorthm) (Rubin, 1987), medtem ko trenutno za najboljšo izbiro veljajo modeli, ki temeljijo na večkratnem vstavljanju (Multiple Imputation models), katerih končni rezultat je ocena, ki je pridobljena na podlagi več različnih vstavljanj (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger,

2001). Njihova poglavitna prednost z vidika statistične analize je ohranjanje variance ter boljša učinkovitost pri višjem odstotku manjkajočih vrednosti, vendar so najbolj kompleksni (Royston, 2004).

Preprost pristop zanemarjanja manjkajočih vrednosti ali zamenjave s povprečnimi vrednostmi se zanaša na predpostavko, da so manjkajoče vrednosti tipa MCAR (Missing Completely at Random – povsem naključno manjkajoči podatki), kar pomeni, da je mehanizem nastanka manjkajočih vrednosti neodvisen od samih manjkajočih vrednosti, kakor tudi drugih atributov (Little & Rubin, 1987). V praksi je ta pogoj le redkokdaj izpolnjen (Little, 1988). Precej verjetneje je, da so manjkajoče vrednosti tipa MAR (Missing at Random – naključno manjkajoči podatki), katerih mehanizem nastanka je, čeprav ime sugerira drugače, odvisen od drugih atributov vzorca (Rubin, 1976). V takšnem primeru je za zadovoljivo aproksimacijo manjkajočih vrednosti potrebno poseči po kompleksnejših metodah, kot so algoritem EM ali regresijski modeli. Najzahtevnejši primer predstavljajo ti. NMAR manjkajoče vrednosti (Not Missing at Random – nenaključno manjkajoči podatki). Mehanizem njihovega nastanka je odvisen od dejanske vrednosti opazovanega atributa. Če delež manjkajočih vrednosti ni dovolj majhen (< 5 %), je za njihovo nadomeščanje kvečjemu primerna metoda, ki temelji na večkratnem vstavljanju (Meng, 1994).

Z namenom zapolnitve vrzeli, ki obstaja med obstoječimi metodami enkratnega vstavljanja manjkajočih vrednosti (single-impute methods) in modeli, ki temeljijo na večkratnem vstavljanju in pri katerih je za vsak cikel vstavljanja potrebna ločena statistična analiza, smo v okviru disertacije razvili nov postopek nadomeščanja manjkajočih vrednosti, ki temelji na principu nadzorovanega strojnega učenja. Za takšen pristop je značilno, da metoda poskuša pridobiti določeno znanje iz množice izbranih učnih vzorcev, ki so sestavljeni iz diskretnih ali zveznih atributov in njim pripadajoče izhodne vrednosti. S pomočjo akumuliranega znanja ustvari funkcijo, ki kot svoj vhod sprejme nov vzorec, po strukturi enak učnim elementom, in napove njegovo izhodno vrednost (Alpaydin, 2004). Kadar je rezultat tega postopka zvezna vrednost, govorimo o regresiji. Pri našem delu smo se osredotočili na to zvrst strojnega učenja, za katero obstajajo številne uveljavljene metode, kot so odločitvena drevesa (Breiman, Friedman, Olshen, & Stone, 1984), metoda k-najbližjih sosedov (Massachusetts Institute of Technology, 2005), nevronske mreže (Gupta & Lam, 1996), metoda podpornih vektorjev (Wang, Li, Jiang, & Feng, 2006) ipd. Kot smo že omenili, se pri konkretnem problemu nadomeščanja manjkajočih vrednosti najbolje izkažejo metode, ki se poslužujejo večkratnega vstavljanja, pri čemer zagotovijo ohranjanje variance med vrednostmi posameznih atributov ob dobri natančnosti napovedanih vrednosti. S tem v vidu smo uporabili ansambelski pristop, pri katerem več regresijskih metod istega tipa prispeva k ustvarjanju čim bolj natančnih napovedanih vrednost, pri čemer se ohrani prednost metod enkratnega

vstavljanja, saj je za vsako manjkajočo vrednost ustvarjena le ena nadomestna, kar odpravi potrebo po dodatnih naknadnih statističnih analizah.

Rezultati številnih novejših raziskav namreč kažejo, da je možno izboljšati natančnost klasifikacije (regresije) posameznih metod strojnega učenja z njihovim združevanjem v ansamble (Kuncheva, 2004). To področje raziskovanja je danes izjemno aktivno in je rodilo že mnoge različne kombinacije klasifikatorjev (Polikar, 2006). Ideja, ki predstavlja osnovo za to kombinacijsko paradigmo, je intuitivna – kako z združevanjem izkoristiti prednosti posameznih metod oziroma njihovih različic, ki se učijo na različnih podmnožicah, saj je skupina močnejša od posameznika. Po svoji uveljavljenosti trenutno izstopata predvsem dva postopka. Prvega, imenovanega »bagging«, je leta 1996 predstavil Breiman (Breiman, 1996), kmalu za njim pa se je pojavil tudi »boosting« (Freund & Schapire, 1996). Pri prvem se posamezni klasifikatorji oz. regresijske metode učijo na različnih podmnožicah iz osnovne učne množice vzorcev, medtem ko drugi postopek pri učenju metod uporablja dodatno obteževanje vzorcev, pri katerih je izhodna napaka večja in se tako osredotoči na težavnejše vzorce. Oba postopka pri klasifikaciji uporabljata večinsko glasovanje in pri regresiji povprečno vrednost izhodnih rezultatov. Izkazalo se je, da za učinkovito delovanje »bagging« potrebuje dovolj velik ansambel metod, saj je drugače raznolikost le-teh premajhna. Z namenom povečanja raznolikosti je Breiman predstavil različico algoritma, poimenovano naključni gozd (Random Forest), ki kot klasifikatorje uporablja izključno odločitvena drevesa (Breiman, 2001). Pri njihovi gradnji se uporabljajo naključno izbrani vzorci in posamezni atributi, kar privede do večje raznolikosti odločitvenih dreves (Liaw & Wiener, 2002). Še korak naprej sta naredila Rodriguez in Kuncheva (Rodríguez, Kuncheva, & Alonso, 2006), ki sta pri konstrukciji metod ansambla uporabila transformacijo prostora po postopku analize osnovnih komponent (PCA- Principal Component Analysis) (Rao, 1964) z namenom zagotavljanja natančnost in raznolikosti posameznih članov ansambla. Kot osnovne klasifikatorje sta uporabila odločitvena drevesa in ansambel poimenovala rotacijski gozd (Rotation Forest).

Avtorja rotacijskega gozda sta opravila obsežno primerjavo njunega ansambla z »bagging« in »boosting« tipi ansamblov na različnih podatkovnih bazah in poročala o signifikantnem izboljšanju natančnosti pri postopku klasifikacije (Rodríguez, Kuncheva, & Alonso, 2006). Dodatno njun pristop omogoča uporabo poljubne osnovne metode, kar se lahko izkaže kot velika prednost, kadar je za reševanje specifičnega problema najbolj primeren točno določen tip klasifikatorja. Prav to možnost smo izkoristili pri reševanju našega problema nadomeščanja manjkajočih vrednosti v bazah podatkov, saj smo kot osnovno metodo ansambla izbrali regresijsko drevo. Naša domneva je bila, da nam bo ta pristop

poleg pričakovane visoke natančnosti lahko omogočil, da s pomočjo dovolj velike raznolikosti osnovnih metod dobro ohranimo tudi varianco.

1.2 Cilji doktorske disertacije

Za nadomeščanje manjkajočih vrednosti v specifični podatkovni bazi je možno razviti algoritem, ki bo bolj uspešen kot katerakoli splošna metoda, tako glede natančnosti kot tudi drugih zahtev, ki jih postavlja statistična analiza (Horton & Kleinman, 2007). Takšen pristop seveda zahteva povsem novo načrtovanje in implementacijo rešitve, kar pa ponavadi ni v interesu analitikov, ki jim manjkajoče vrednosti sicer predstavljajo problem, vendar niso v primarnem fokusu njihovega dela. Splošna, učinkovita in že obstoječa rešitev je ponavadi prva izbira. V množici obstoječih metod za nadomeščanje manjkajočih vrednosti je veliko preprostih, ki se obnesejo izjemno slabo, kadar je odstotek manjkajočih vrednosti nezanemarljiv oziroma te niso tipa MCAR (Scheffer, 2002). Naprednejše metode so sicer relativno natančne, vendar slabo ohranjajo varianco podatkov ali zahtevajo izpolnjene določene predpogoje (podatki tipa MAR, ustrezna porazdelitev podatkov ipd.) (Horton, Lipsitz, & Parzen, 2003).

Cilj doktorske naloge je bil razviti metodo za nadomeščanje manjkajočih vrednosti, ki temelji na ansamblu regresijskih dreves, imenovanem »rotacijski regresijski gozd«. Predvidevali smo, da se bo ansambel, ki se je v podobni konfiguraciji že dobro obnesel na področju klasifikacije, izkazal z visoko natančnostjo tudi pri reševanju regresijskega problema aproksimacije manjkajočih vrednosti. Za to naj bi poskrbelo zadostno število med seboj raznolikih, a obenem natančnih, individualno zgrajenih regresijskih dreves. Raznolikost posameznih osnovnih dreves smo uporabili za dodatno izboljšanje ohranjanja variance, ki odraža negotovost določitve manjkajočih vrednosti. Želeli smo doseči signifikantno izboljšanje natančnosti določanja in ohraniti visoko variabilnost podatkov. Učinkovitost rotacijskega gozda smo demonstrirali s primerjavo z uveljavljenimi metodami za nadomeščanje manjkajočih vrednosti.

1.2.1 Teze doktorske disertacije

V doktorski disertaciji želimo potrditi naslednjo tezo:

Z uporabo rotacijskega regresijskega gozda kot ansambla regresijskih dreves lahko v primerjavi s klasičnimi metodami za nadomeščanje manjkajočih vrednosti v podatkovnih bazah dosežemo večjo natančnost pri določanju manjkajočih vrednosti ob zagotavljanju ohranjanja variance podatkov.

Iz podane teze smo izpeljali naslednje hipoteze:

Hipoteza 1:

Z uporabo nove metode, zasnovane na ansamblu raznolikih, a obenem dovolj natančnih regresijskih dreves, se v večini primerov izboljša natančnost določanja manjkajočih vrednosti v primerjavi s klasičnimi metodami, kakor tudi v primerjavi z individualnim regresijskim drevesom.

Hipoteza 2:

Uporaba rotacijskega regresijskega gozda za aproksimacijo manjkajočih vrednosti v večini podatkovnih baz ohranja boljšo stopnjo variance kot klasične »single-impute« metode.

Hipoteza 3:

Z uvedbo stohastične metode, ki upošteva zanesljivosti predikcij posameznih manjkajočih vrednosti, lahko dodatno izboljšamo ohranjanje variance ob zanemarljivem vplivu na natančnost določanja manjkajočih vrednosti.

1.2.2 Metode dela

Problematika manjkajočih vrednosti se rešuje na različne načine, ki jih v grobem lahko razdelimo na ignoriranje, brisanje ali nadomeščanje. Mnogokrat v poštev pride le slednja, najbolj kompleksna rešitev. Zanjo obstajajo številne metode, izmed katerih smo izbrali najbolj uveljavljene (Horton & Kleinman, 2007), (Brock, Shaffer, Blakesley, Lotz, & Tseng, 2008), ki smo jih vključili v našo analizo. Osredotočili smo se predvsem na natančnost in ohranjanje variance podatkov, ki je lahko ključnega pomena za primerno nadaljnjo analizo. Z namenom izboljšanja učinkovitosti obstoječih metod smo razvili novo metodo, zasnovano na ansamblu regresijskih dreves, imenovanem rotacijski gozd. Ta ansambel temelji na gradnji raznolikih in natančnih osnovnih klasifikatorjev (oziroma regresorjev) in se je že izkazal kot učinkovito orodje za klasifikacijo. Odločitvena drevesa, ki jih uporablja v svoji osnovi, smo

nadomestili z regresijskimi drevesi in izvedli implementacijo v uveljavljenem odprtokodnem ogrodju Weka (Witten & Frank, 2005). Za evaluacijo metode smo uporabili podatkovne baze iz javno dostopne zbirke UCI Machine Laerning Repository (Asuncion & Newman, 2007) in še dve dodatni podatkovni bazi (Bohen, Troyanskaya, Alter, Warnke, Botstein, & Brown, 2003), (Spellman, in drugi, 1998), zaradi primerjave z rezultati že objavljene študije (Brock, Shaffer, Blakesley, Lotz, & Tseng, 2008). Odvisno od želenega mehanizma nastanka manjkajočih vrednosti (MCAR, MAR, NMAR) smo na različne načine simulirali nastanek manjkajočih vrednosti v sedmih različnih stopnjah (1%, 5%, 10%, 15%, 20%, 25% in 50%). Na tako modificiranih bazah smo izvedli večkratne ponovitve validacij vseh izbranih metod, vključno z našo. Za oceno natančnosti smo se poslužili klasične metrike RMSE (»root mean-squared error«, koren povprečne kvadratne napake) ter ugotavljali, kako metoda vpliva na varianco in povprečno vrednost posameznih atributov. S pomočjo empiričnega znanstvenega pristopa na osnovi analize literature in primerjave z obstoječimi metodami smo ovrednotili našo metodo ter preverili postavljene hipoteze z univariatno statistično analizo v orodju za statistično analizo SPSS.

1.2.3 Pričakovani izvirni znanstveni prispevki

V doktorskem delu smo razvili novo metodo za nadomeščanje manjkajočih vrednosti v podatkovnih bazah.

Izvirni znanstveni prispevki doktorskega dela so:

- uporaba rotacijskega gozda kot ansambla regresijskih dreves za implementacijo metode nadomeščanja manjkajočih vrednosti v podatkovnih bazah,
- uvedba stohastične metode, temelječe na zanesljivosti predikcije regresijskih dreves, za izboljšanje ohranjanja variance podatkov,
- uporabniku prijazno okolje za izvedbo nadomeščanja manjkajočih vrednosti, neodvisno od podatkovne baze,
- integracija metode v orodja za podatkovno rudarjenje,
- pridobivanje novega znanja z uporabo izpopolnjenega orodja za podatkovno rudarjenje:
    - podatkovno rudarjenje smo opravili na podatkovni bazi, pridobljeni z analizo termografskih slik rok pacientov s sindromom karpalnega kanala, kjer imamo opravka z manjkajočimi vrednostmi, ki jih ne moremo zanemariti.

1.2.4 Predpostavke in omejitve

Pri razvoju metode za nadomeščanje manjkajočih vrednosti smo se omejili na zvezne manjkajoče vrednosti, kljub temu, da bi lahko implementirali tudi ansambel odločitvenih dreves ali opravili ustrezne preslikave diskretnih razredov. Predpostavili smo, da bo metoda uspešno delovala na podatkih tipa MCAR in MAR ter bo precej neuspešna pri podatkih tipa NMAR. Skladno z ugotovitvami avtorjev rotacijskega gozda metoda ne bo primerna za izvajanje na izjemno obsežnih podatkovnih bazah, katerih dimenzije se merijo v milijonih elementov in/ali atributov. Pri ovrednotenju učinkovitosti naše metode smo se omejili predvsem na podatkovne baze iz javno dostopne zbirke Machine Learning Repository univerze UCI (Asuncion & Newman, 2007).

1.3 Struktura disertacije

Uvodnemu poglavju doktorske disertacije sledi poglavje, ki opisuje osnove strojnega učenja s poudarkom na nadzorovanem učenju, regresijskih metodah in združevanju osnovnih metod v ansamble. Tretje poglavje obravnava problematiko manjkajočih vrednosti, vzroke njihovega nastanka in njihove posledice ter načine ravnanja z manjkajočimi vrednostmi. V četrtem poglavju je podrobno opisan rotacijski gozd in njegova različica, ki smo jo razvili za rešitev našega problema. Predstavljena je tudi naša metoda za izboljšanje ohranjanja variance. Peto poglavje obravnava način implementacije razvitih metod ter zasnovo eksperimentalnega okolja za preverjanje postavljenih hipotez. V najobsežnejšem, šestem poglavju so zbrani rezultati primerjave učinkovitosti posameznih metod za nadomeščanje manjkajočih vrednosti na različnih bazah glede na različne mehanizme nastanka manjkajočih vrednosti. Predstavljena je natančnost ter ohranjanje variance pri posameznih metodah. Temu poglavju sledi razprava in v zaključku povzetek opravljenega dela z idejami za možne izboljšave in nadaljnje delo.

2 Nadzorovano učenje in regresija

V tem poglavju je podan osnoven pregled področja nadzorovanega učenja, ki predstavlja eno poglavitnih vej strojnega učenja. Le-to se ukvarja z metodami, ki uporabljajo izkušnje za izboljšanje svoje zmogljivosti, kar je razumljivo, če upoštevamo sledeči definiciji učenja (Kononenko, 1997):

V splošnem je učeči se stroj vsaka naprava, pri kateri izkušnje iz preteklosti vplivajo na akcije.

Nils J. Nilsson

Učenje določa adaptivne spremembe v sistemu, ki mu omogočajo, da naslednjič reši nalogo iste vrste bolj učinkovito.

Herbert A. Simon

2.1 Algoritmi strojnega učenja

Če vzamemo v obzir današnjo prežetost sveta s podatki, katerih obseg narašča iz dneva v dan zahvaljujoč preprostosti in cenenosti shranjevanja vseh potrebnih in nepotrebnih informacij znotraj vseprisotnih računalniških sistemov (Witten & Frank, 2005), ni težko dojeti, da se v tej množici podatkov skrivajo uporabne informacije, ki pa pogosto ostajajo skrite. Prav pri iskanju teh skritih vzorcev, ki mu s strokovnim izrazom pravimo podatkovno rudarjenje, se lahko izkaže vsa moč in praktičnost strojnega učenja. Če opredelimo področje uporabe bolj natančno, sodijo med tipične aplikacije strojnega učenja sintaktično razpoznavanje vzorcev, računalniški vid in sluh, spletni iskalniki, razpoznavanje pisave in govora, umetna inteligenca v računalniških igrah, obdelava naravnega jezika, medicinska diagnostika, bioinformatika, napovedovanje borznih tečajev in še bi lahko naštevali.

Glede na podatke, ki so na voljo in cilj strojnega učenja, so na voljo številni različni algoritmi, ki jih lahko razdelimo v več podskupin:

- **Nadzorovano učenje (*supervised learning*)** – algoritem na podlagi množice označenih učnih vzorcev, sestavljenih iz parov vhodnih in izhodnih podatkov, ustvari funkcijo, katere naloga je, da vhodne podatke algoritma preslika v izhodne vrednosti. Namen učenja je, da

ustvarjena funkcija zna čim bolj pravilno napovedati izhodne vrednosti (razred) za poljubne vhodne vrednosti.

- **Nenadzorovano učenje (*unsupervised learning*)** – pri tej skupini algoritmov učenje poteka tako, da izhodna vrednost tudi med učenjem ni znana. Cilj učenja tako ni predikcija izhodne vrednosti ampak ugotavljanje strukture oz. urejenosti vhodnih podatkov. Tipični primeri takšnega učenja so grupiranje (*clustering*), analiza neodvisnih komponent (*independent component analysis*), Kohonenove nevronske mreže itn.
- **Delno nadzorovano učenje (*semi-supervised learning)*** – algoritem pri učenju uporablja tako označene (znana izhodna vrednost oz. razred) kot neoznačene vzorce.
- **Učenje z ojačitvijo (*reinforcement learning*)** – algoritem se uči glede na odziv iz okolja, ki posredno ojača pravilno akcijo algoritma.
- **Transdukcija (*transduction*)** – postopek učenja je podoben kot pri nadzorovanem učenju, le da se pri transdukciji med učenjem uporabljajo tudi vzorci iz testne množice.

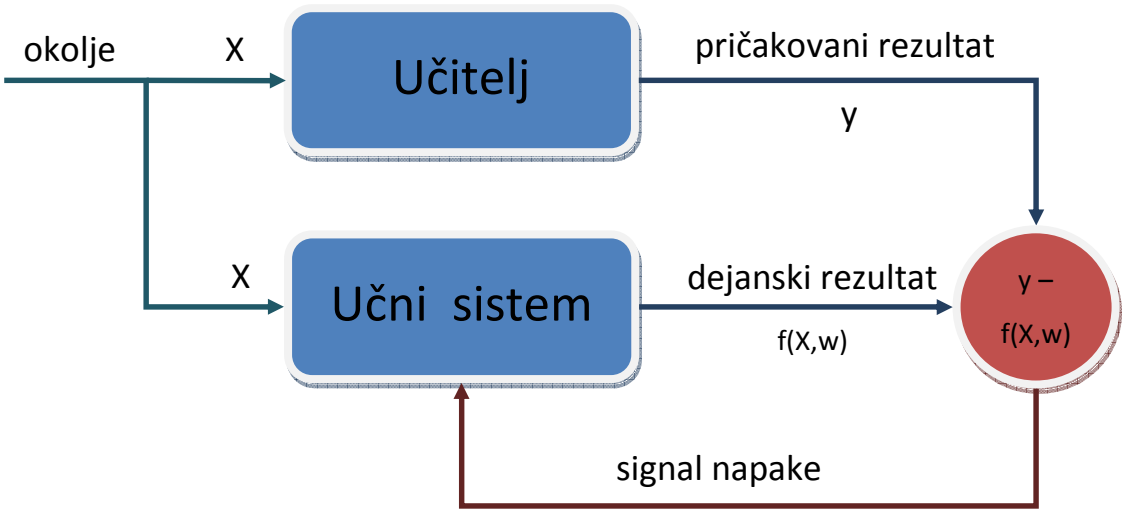
Večina najbolj pogosto uporabljanih algoritmov sodi v prvi dve izmed naštetih podskupin (Kantardzic, 2003). V nadaljevanju se bomo osredotočili na nadzorovano učenje, v katero lahko prištejemo tudi našo metodo za nadomeščanje manjkajočih vrednosti.

2.2 Nadzorovano učenje

Kot je že bilo omenjeno, se nadzorovano učenje uporablja za določanje predhodno neznane odvisnosti med vhodnimi in izhodnimi vrednostmi vzorcev dane učne množice. Pri tem postopek učenja domneva, da obstaja funkcija, ki lahko skozi postopno prilagajanje znotraj učnega procesa postane sposobna generalizacije iz učnih na testne podatke.

Na sliki 2.2-1 je prikazan diagram, ki predstavlja takšno obliko učenja. »Učitelj« poseduje znanje o okolju, pri čemer je to znanje množica vzorcev, sestavljenih iz vhodnih in njim pripadajočih izhodnih vrednosti. Po drugi strani pa je okolje s svojimi lastnostmi in modelom, ki ga opisuje, za »učni sistem« neznano. Parametri učnega procesa (w) se med postopkom učenja spreminjajo pod skupnim vplivom vhodnih podatkov in signala napake, ki je definiran kot razlika med pričakovano izhodno vrednostjo oz. rezultatom, ki ga pozna učitelj, in dejanskim rezultatom, ki ga kot svoj izhod ustvari učni sistem. Takšen sistem se torej nahaja v zaprti zanki, pri čemer je potrebno upoštevati, da se neznano okolje v njej ne nahaja. Z ustrezno množico učnih vzorcev je možno zmanjšati signal napake do te mere, da je

nadzorovani učni sistem sposoben opravljati naloge, kot so klasifikacija in napovedovanje funkcijskih vrednosti.



Slika 2.2-1: Diagram nadzorovanega strojnega učenja

Nadzorovano učenje je torej tehnika strojnega učenja, pri kateri se skozi učni proces ustvari funkcija, ki je sposobna preslikati vhodne podatke v izhodno vrednost oz. rezultat na podlagi znanja, pridobljenega iz učnih podatkov. Učni podatki so sestavljeni iz parov vhodnih objektov in pripadajočih izhodnih vrednosti. Izhod naučene funkcije je lahko zvezna vrednost (takrat govorimo o regresiji) ali diskretna vrednost, ki predstavlja razred, kateremu pripada obravnavani objekt (klasifikacija). Naloga nadzorovanega učnega sistema je napovedati vrednost funkcije za vsak veljavni vhodni objekt, za kar je potrebna sposobnost sistema, da generalizira podatke iz učne množice na nove, pred tem neznane objekte.

Poznamo veliko število algoritmov nadzorovanega strojnega učenja, ki jih glede na tip izhodne vrednosti ločimo na klasifikatorje (diskretna izhodna vrednost) in regresorje (zvezna izhodna vrednost). Vsak izmed teh klasifikatorjev oz. regresorjev ima svoje prednosti in slabosti. Njihova učinkovitost je v veliki meri odvisna od lastnosti podatkov, na katerih se učijo. Ne poznamo univerzalnega najboljšega algoritma, ki bi se izkazal na vseh možnih množicah podatkov. Wolpert je že leta 1996 izpeljal več teoremov, ki jih še najbolje poznamo pod originalnim imenom »*no free lunch theorems*« in katerih

skupna trditev se lahko povzame kot: algoritem strojnega učenja, ki bi se najbolje izkazal na vseh možnih problemih, ne obstaja (Wolpert, 1996).

Nekatere izmed najbolj pogosto uporabljanih metod nadzorovanega strojnega učenja so:

- nevronske mreže (*neural networks*),
- odločitvena drevesa (*decision trees*),
- podporni vektorji (*support vector machines*),
- metoda najbližjih sosedov (*k-nearest neighbours*),
- linearna regresija (*linear regression*),
- naivni Bayes (*naive Bayes*),
- ansambli klasifikatorjev oz. regresorjev,
- itd.

2.2.1 Predstavitev podatkov

Različni problemi, ki jih poskušamo reševati na področju strojnega učenja, se lahko razvrstijo v različne koncepte, ki se ločijo po načinu učenja (Witten & Frank, 2005). Tako razlikujemo med npr. grupiranjem, asociativnim učenjem, klasifikacijo in numerično predikcijo, ki predstavlja koncept, ki je za naš problem nadomeščanja manjkajočih vrednosti v podatkovnih bazah najbolj zanimiv. Takšen koncept učenja zahteva uporabo regresorja, saj je namen učnega sistema napovedovanje (zveznih) numeričnih vrednosti. Različni koncepti zahtevajo različne učne vzorce. Medtem ko morajo klasifikatorji in regresorji med postopkom učenja poznati pripadajoče izhodne vrednosti učnih vzorcev, temu ni tako pri grupiranju in asociativnem učenju, kjer je namen učenja iskanje podobnosti oz. »zanimivih« podatkovnih struktur.

Vhodne podatke v vsak sistem strojnega učenja predstavlja množica vzorcev. Ti vzorci so objekti, ki jih želimo klasificirati, povezati, grupirati ali jim dodeliti pripadajoče numerične vrednosti. Vsak vzorec predstavlja individualen, neodvisen primerek koncepta, ki se ga želi sistem naučiti. Obenem je vsak predstavljen z vrednostmi posameznih atributov, ki skupaj določajo podatkovno strukturo celotne množice vzorcev. Takšno podatkovno zbirko si lahko predstavljamo kot matriko neodvisnih vzorcev in njihovih atributov (Tabela 2.2-1).

Dolžina čašnega lista (cm)	Širina čašnega lista (cm)	Dolžina venčnega lista (cm)	Širina venčnega lista (cm)	Razred (tip perunike)
5,1	3,5	1,4	0,2	**Iris setosa**
4,9	3,0	1,4	0,2	**Iris setosa**
7,0	3,2	4,7	1,4	**Iris versicolor**
6,4	3,2	4,5	1,5	**Iris versicolor**
6,3	3,3	6,0	2,5	**Iris virginica**
5,8	2,7	5,1	1,9	**Iris virginica**

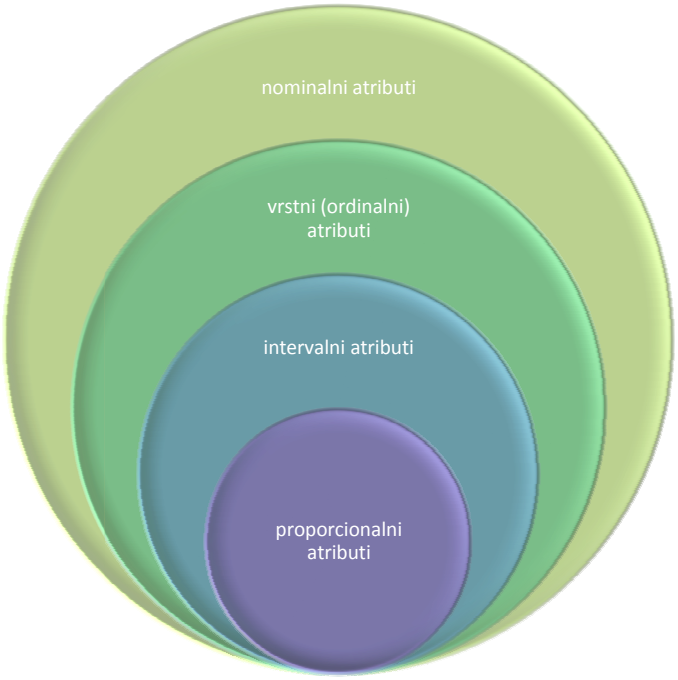
Tabela 2.2-1: Primer majhne množice vzorcev klasičnega problema klasifikacije perunik (R.A. Fisher 1935)

2.2.2 Razdelitev atributov

Atributi so vnaprej izbrane lastnosti, ki opisujejo posamezne vzorce podatkovne množice. V tabeli se vrednosti posameznih atributov nahajajo v pripadajočih stolpcih, medtem ko vsaka vrstica tabele predstavlja nov vzorec. Tipični problemi, ki jih rešujemo s pomočjo nadzorovanega strojnega učenja, zahtevajo obstoj odločitvenega atributa, ki predstavlja razred, kateremu posamezni vzorec pripada, ali numerično vrednost, po kateri se vzorec razlikuje od ostalih vzorcev. Vsak vzorec pripada natanko enemu razredu iz množice vseh možnih razredov odločitvenega atributa. Cilj nadzorovanega strojnega učenja je, da se sistem čim bolj zanesljivo nauči razporejati nove vzorce v njim pripadajoče razrede ali jim dodeljevati ustrezne numerične vrednosti.

Vrednost določenega atributa posameznega vzorca je ovrednotenje lastnosti vzorca, na katero se atribut nanaša. V grobem lahko razdelimo atribute v dve skupini: **numerične** in **kategorične**. Numerični atributi, ki jim pravimo tudi kvantitativni, so izmerjene celoštevilske ali realne vrednosti, medtem ko kategorični atribut lahko zavzame vrednosti le iz vnaprej določenega končnega nabora opisnih vrednosti.

Na področju statistike je v veljavi nekoliko drugačna razdelitev atributov, ki jih razdeli v štiri skupine glede na način razlikovanja med vrednostmi posameznega atributa. Tako ločimo *nominalne*, *vrstne (ordinalne)*, *intervalne* in *proporcionalne* atribute, pri čemer lahko vsako skupino dojemamo kot podmnožico prejšnje (Slika 2.2-2).



Slika 2.2-2: Delitev atributov glede na možnost razvrščanja njihovih vrednosti

Vrednosti nominalnih atributov so opisni simboli, ki služijo le kot imena. Posameznih vrednosti atributa ni možno razvrstiti v urejeno zaporedje, tako da bi med njimi obstajala relacija razdalje ali velikosti.

Vrstni atributi se od nominalnih razlikujejo po tem, da jih lahko uredimo po vrsti, ne moremo pa izračunati razdalje med njimi. Npr. vrednosti atributa »*temperatura*« so lahko *vroče*, *hladno* in *toplo*, ki jih lahko razvrstimo:

vroče > *toplo* > *hladno* ali *vroče* < *toplo* < *hladno*,

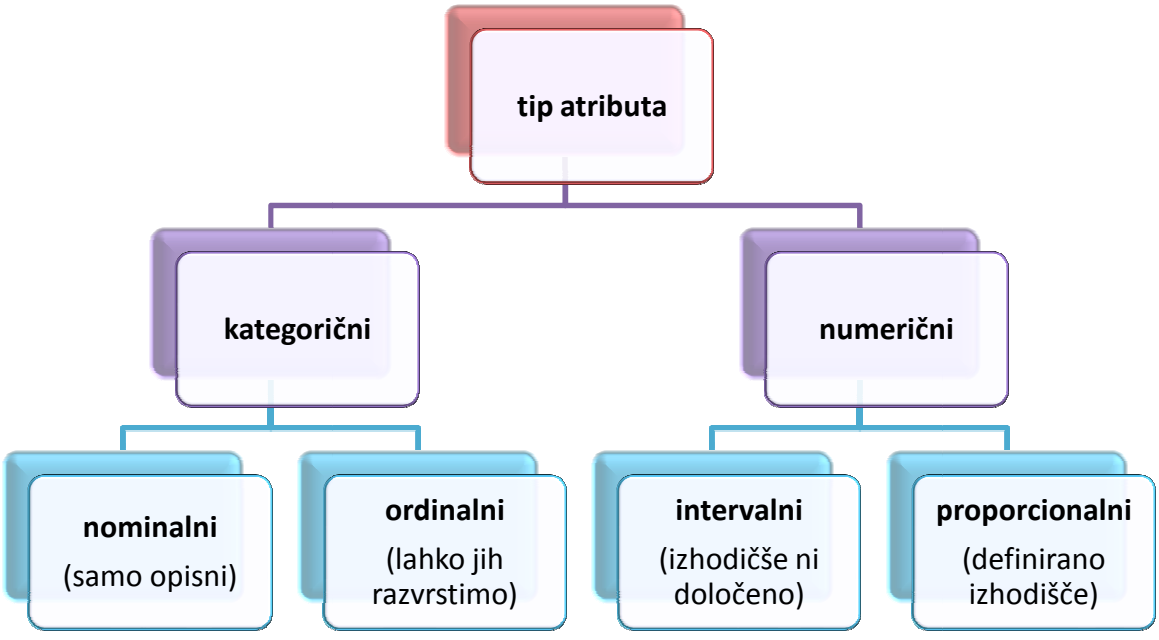
pri čemer ni pomembno, katera izmed razvrstitev je bolj intuitivna – pomembno je, da se vrednost *toplo* nahaja med vrednostma *hladno* in *vroče*. Čeprav je vrednosti vrstnega atributa smiselno primerjati med seboj, jih ne moremo seštevati ali odštevati – npr. ne moremo ovrednotiti, ali je razlika med *hladno* in *toplo* večja kot razlika med *toplo* in *vroče.*

Naslednjo skupino atributov predstavljajo intervalni atributi, katerih vrednosti lahko razvrščamo in tudi izmerimo v fiksnih, enakih enotah. Dober primer takšnega atributa je temperatura, izmerjena v stopinjah Celzija, in ne z opisno lestvico z vrednostmi hladno, toplo, vroče. V takšnem primeru je

smiselno govoriti o temperaturnih razlikah, saj lahko posamezne vrednosti odštevamo ter primerjamo z drugimi razlikami. Drugi primer intervalnega atributa predstavljajo letnice. Tudi tukaj lahko govorimo o razlikah med letnicami (npr. razlika med leti 2009 in 1996 je 13 let), medtem ko je seštevek dveh letnic dokaj nesmiseln, saj je izhodišče (leto 0) naključno izbrano.

Proporcionalni atributi so tisti, pri katerih je izhodišče (točka 0) znano oz. točno določeno. Tipičen primer proporcionalnega atributa je razdalja med dvema objektoma. Oddaljenost objekta od samega sebe je namreč 0. Vrednosti proporcionalnih atributov lahko obravnavamo kot realna števila, nad katerimi so dovoljene vse matematične operacije.

Če se vrnemo k prvotni delitvi atributov na numerične in kategorične, jo lahko sedaj razširimo na podstopnje. Kategorične atribute lahko razdelimo na nominalne in ordinalne, numerične pa na intervalne in proporcionalne (Slika 2.2-3). Mnogokrat lahko zasledimo izraz *diskretne vrednosti*, ki se uporablja tako za intervalne (celoštevilske) kot ordinalne atribute, redkeje tudi za nominalne (vse kategorične). To je posledica postopka diskretizacije, pri kateri se zvezna numerična količina preslika v diskretno vrednost, kar ohrani urejenost posameznih vrednosti.



Slika 2.2-3: Razvrstitev atributov glede na tip

Poleg atributov, ki določajo vsak vzorec v podatkovni množici, lahko sistem strojnega učenja pri učnem procesu uporablja tudi podatke o podatkih, ti. *meta-podatke*. Te dodatne informacije lahko sistem pridobi s predhodno analizo podatkovne množice, vendar praktične metode, ki jih bomo predstavili v naslednjih poglavjih, operirajo zgolj na vrednostih atributov.

2.2.3 Format ARFF

Kakovostna priprava vhodnih podatkov, ki je predpogoj za uspešno izvajanje učnega procesa, zahteva veliko pozornosti in časa. Zbiranje in združevanje podatkov v vzorce, pravilna izbira atributov, ki so relevantni za ciljno analizo, končno čiščenje zbirke (eliminacija šuma, nizkokakovostnih vzorcev, podvojenih vzorcev) so le nekateri izmed korakov pri izdelavi podatkovne množice.

Tipičen način organizacije podatkov znotraj podatkovne množice predstavlja tekstovna datoteka, v kateri posamezen vzorec zavzema eno vrstico, ki se sestoji iz vrednosti atributov, ločenih z vejicami ali kakšnim drugim znakom. Pri našem delu smo uporabili javno dostopne podatkovne zbirke, ki so bile v večini primerov shranjene v takšnem formatu. Zaradi lažjega dela z odprtokodnim orodjem Weka (Witten & Frank, 2005), smo podatkovne zbirke pretvorili v format, ki je bil razvit posebej za to orodje in se imenuje ARFF (Attribute-Relation File Format). Podobno kot navadna tekstovna datoteka se tudi ta format uporablja za shranjevanje zbirk, sestavljenih iz neodvisnih in nerazvrščenih vzorcev, pri čemer razločuje med 4 različnimi tipi atributov. Poleg nominalnih in numeričnih atributov pozna tudi atribut tipa »datum« in nize znakov (*string*) za shranjevanje dolgih besedil, na katerih se lahko izvaja tekstovno rudarjenje. V glavi *arff* datoteke se nahaja opis relacije, ki jo predstavljajo atributi, ter seznam vseh atributov. Vsakemu nominalnemu atributu je znotraj zavitih oklepajev dodana množica dovoljenih vrednosti, medtem ko je numeričnim atributom dodana ključna beseda »numeric«. Glavi sledi ključna beseda »@data«, ki označuje začetek shranjenih vzorcev, pri čemer je vsak vzorec predstavljen z vrednostmi atributov, ločenimi z vejicami, v vrstnem redu, kot so navedeni v glavi (Slika 2.2-4). Dodatna posebnost tega formata je možnost racionalnega shranjevanja podatkov, predstavljenih z redko matriko (*sparse matrix*). Za takšne podatke je značilno, da vsebujejo izjemno veliko število ničel, zato je smiselno shraniti samo vrednosti, ki se razlikujejo od 0. Vsaka neničelna vrednost posameznega vzorca je navedena v paru z zaporedno številko atributa, ki mu vrednost pripada. Pari so ločeni z vejicami, medtem ko se celoten vzorec (vrstica) nahaja znotraj zavitih oklepajev.

```
% 1. Title: Iris Plants Database
%
% 2. Sources:

%     (a) Creator: R.A. Fisher
%     (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%     (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class       {Iris-setosa, Iris-versicolor, Iris-virginica}

@DATA
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
4.7, 3.2, 1.3, 0.2, Iris-setosa
4.6, 3.1, 1.5, 0.2, Iris-setosa
5.0, 3.6, 1.4, 0.2, Iris-setosa
5.4, 3.9, 1.7, 0.4, Iris-setosa
4.6, 3.4, 1.4, 0.3, Iris-setosa
5.0, 3.4, 1.5, 0.2, Iris-setosa
4.4, 2.9, 1.4, 0.2, Iris-setosa
4.9, 3.1, 1.5, 0.1, Iris-setosa
```

Slika 2.2-4: Primer datoteke v formatu ARFF

2.3 Načini evalvacije rezultatov

Kakovostna evalvacija ali ovrednotenje rezultatov je ključnega pomena, če želimo za reševanje določenega problema izbrati najbolj ustrezno metodo. Preverjanje učinkovitosti metod ni tako preprosto, kot se morda zdi na prvi pogled. Namreč ne zadostuje, če natančnost preverjamo na učni množici, če ta ne predstavlja dovolj dobrega približka neodvisne testne množice.

Težavam se najlaže izognemo, če imamo na voljo zelo veliko število vzorcev. Takrat lahko ustvarimo model na veliki učni množici in ga nato preverimo na drugi, prav tako veliki, testni množici. Vendar so takšni primeri bolj izjema kot pravilo. Ponavadi so kakovostni podatki redki, zato imamo na

voljo kar nekaj načinov napovedovanja učinkovitosti metod strojnega učenja na podlagi omejenih podatkovnih virov. Izmed njih se je najbolj uveljavila metoda prečnega preverjanja (cross-validation), ki jo bomo predstavili v nadaljevanju. Primerjanje različnih metod na danem problemu zahteva uporabo statističnih testov, s katerimi zagotovimo, da naključni dejavniki niso vzrok za razlike v učinkovitosti. Klasifikacijski problemi zahtevajo drugačen pristop kot regresijski, kjer ne merimo deležev pravilno klasificiranih vzorcev temveč napako numerične predikcije.

2.3.1 Prečno preverjanje

Predpostavimo, da imamo na voljo podatkovno množico *M* z omejenim številom vzorcev *n*, na kateri želimo naučiti klasifikator *K*. Če uporabimo vseh *n* vzorcev za učenje in nato za preverjanje natančnosti, lahko izračunamo klasifikacijsko napako *e* kot razmerje:

$$e(K) = \frac{n_e}{n}, \tag{2.1}$$

pri čemer je n_e število napačno klasificiranih vzorcev. Oceno natančnosti klasifikacije dobimo z izračunom *1 - e(K)*. Seveda je takšna ocena natančnosti klasifikatorja nerealna, saj se izvaja kar na učni množici, za katero se je klasifikator lahko prenaučil. Preprosta rešitev tega problema je ti. *metoda pridržanja (holdout)*, pri kateri v določenem razmerju razdelimo začetno množico vzorcev na učno in testno množico. V praksi se za testno množico ponavadi pridrži tretjina ali četrtina vzorcev. Slabost te metode je, da z naključno razdelitvijo vzorcev ne moremo zagotoviti, da bosta učna in/ali testna množica dovolj reprezentančni. Temu se lahko v določeni meri izognemo s postopkom stratifikacije, ki zagotovi, da se v obeh množicah ohrani začetno razmerje vzorcev glede na odločitveni atribut. Kljub tej izboljšavi je pogosto težko preprečiti nesorazmerno sestavljenost učnih in testnih množic, zato je smiselno postopek razdelitve osnovne množice večkrat ponoviti, pri čemer se vzorci naključno razvrstijo v učno in testno množico, vendar vedno v istem razmerju. Napaka klasifikatorja se izračuna kot povprečna napaka vseh ponovitev.

Velik korak naprej predstavlja *prečno preverjanje (cross-validation)*. Ideja na kateri sloni je, da se nesorazmernosti učne in testne množice lahko izognemo tako, da enkrat uporabimo eno izmed množic za učenje in drugo za preverjanje ter nato njuni vlogi zamenjamo in izračunamo povprečno napako. Seveda je to možno le ob razdelitvi vzorcev v razmerju 1:1, kar pa ni idealno, saj je bolje uporabiti večji delež za učno množico. Pri prečnem preverjanju lahko določimo število razdelitev osnovne množice,

tako da se ta razdeli na izbrano število približno enako velikih delov. Nato se za vsakega izmed teh delov ponovi postopek preverjanja, kjer se izbrani del uporabi kot testna množica, medtem ko se preostali vzorci izkoristijo za učenje klasifikatorja. Ponavadi se pri razdelitvi osnovne množice za dodatno odpravljanje nesorazmernosti med posameznimi podmnožicami uporabi tudi stratifikacija. Najpogosteje se uporablja razdelitev na 10 podmnožic; takrat govorimo o 10-kratnem prečnem preverjanju. Včasih tudi stratificirano 10-kratno prečno preverjanje ne zadostuje za pridobitev zanesljive ocene klasifikacijske napake. V takšnem primeru je možno celoten postopek večkrat ponoviti, saj so razdelitve osnovne množice naključne, tako da različna 10-kratna prečna preverjanja lahko dajo različne rezultate. Klasifikacijsko napako izračunamo kot povprečje ocen napak, pridobljenih z različnimi preverjanji.

Omeniti velja še posebno različico prečnega preverjanja, kjer je število podmnožic kar enako številu vzorcev. Tako je vsak izmed vzorcev klasificiran natančno enkrat in ni potrebe po večkratnih ponovitvah postopka, saj je celotna procedura deterministična (ni naključnih razdelitev). Hkrati se pri vsakem preverjanju za učenje uporabi največje možno število vzorcev (vsi razen enega). Pomanjkljivost te različice je, poleg visoke računske zahtevnosti, da stratifikacija ni možna.

2.3.2 Evalvacija numerične napovedi

Vse metode ocenjevanja natančnosti, ki smo jih do sedaj opisali, se nanašajo na klasifikacijske probleme in kot take niso primerne v situacijah, ko imamo opravka z regresorjem, katerega naloga je čim bolj natančno napovedati numerične vrednosti. Osnovni principi, kot so uporaba ločene testne množice, metoda pridržanja in prečno preverjanje, prav tako veljajo pri numeričnih napovedih. Razlika je v načinu izračuna same ocene, saj preprosto računanje odstotka pravilnih klasifikacij pri regresiji ne pride v poštev. Napake tu niso zgolj prisotne ali odsotne, izmerimo jim lahko tudi velikost.

Ocena napake regresorja je za nas še posebej zanimiva, saj predstavlja osnovo, s pomočjo katere lahko primerjamo našo metodo za nadomeščanje manjkajočih vrednosti z že obstoječimi. Na voljo imamo več metrik, predstavili bomo najbolj pogosto uporabljane. Pri tem bo v formulah uporabljeno enotno označevanje: napovedane (predvidene) vrednosti bodo označene s $p_1, p_2, \ldots, p_n$, dejanske vrednosti pa z $a_1, a_2, \ldots, a_n$.

Povprečna kvadratna napaka

Povprečna kvadratna napaka (*mean-squared error*) je osnovna in najbolj pogosto uporabljana metrika. Uporabljajo jo mnoge matematične metode (kot je npr. linearna regresija).

$$MSE = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n} \tag{2.2}$$

Koren povprečne kvadratne napake

Kvadratni koren povprečne kvadratne napake (*root mean-squared error*) se uporablja namesto povprečne kvadratne napake, kadar želimo ohraniti oceno v istem velikostnem razredu, kot so tudi napovedane vrednosti.

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \tag{2.3}$$

Povprečna absolutna napaka

Povprečna absolutna napaka (*mean absolute error*) predstavlja alternativo povprečni kvadratni napaki. Je manj občutljiva na obrobne (ekstremne) vrednosti, saj posameznih napak ne kvadrira.

$$MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \tag{2.4}$$

Relativna kvadratna napaka

Relativna kvadratna napaka (*relative squared error*) je metrika, ki relativizira napake glede na napake, ki bi jih ustvarila preprosta metoda za napovedovanje vrednosti (takšna, ki bi kot rezultat dala kar povprečje dejanskih vrednosti iz učne množice). Torej metrika normalizira skupno kvadratno napako, tako da jo deli s skupno napako te preproste metode.

$$RSE = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \cdots + (a_n - \overline{a})^2}, \tag{2.5}$$

kjer je

$$\overline{a} = \frac{1}{n}\sum_i a_i.$$

Relativna absolutna napaka

Podobno kot relativna kvadratna napaka tudi relativna absolutna napaka (*relative absolute error*) relativizira napake glede na napake omenjene preproste metode. Namesto skupne kvadratne napake normalizira skupno absolutno napako.

$$RAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \overline{a}| + \cdots + |a_n - \overline{a}|} \tag{2.6}$$

Koeficient korelacije

Koeficient korelacije (*correlation coefficient*) je metrika, ki se razlikuje od preostalih, saj meri statistično korelacijo med dejanskimi in napovedanimi vrednostmi. Na podlagi ocene te metrike ne moremo ugotoviti, ali so napovedane vrednosti v istem velikostnem razredu kot dejanske, vendar lahko sklepamo, kako dobro z njimi korelirajo. Ocena metrike se giblje od 1 za popolnoma soodvisne vrednosti, preko 0 za neodvisne, do -1 za popolnoma negativno soodvisne (ocena spodobnih regresijskih metod mora biti pozitivna vrednost).

$$CC = \frac{S_{PA}}{\sqrt{S_P S_A}}, \tag{2.7}$$

kjer je

$$S_{PA} = \frac{\sum_i (p_i - \overline{p})(a_i - \overline{a})}{n-1}, \qquad S_P = \frac{\sum_i (p_i - \overline{p})^2}{n-1}, \qquad S_A = \frac{\sum_i (a_i - \overline{a})^2}{n-1}.$$

Katera izmed opisanih metrik je najbolj primerna za ocenjevanje natančnosti v določeni situaciji, je odločitev, ki pogosto ni očitna. Metrike kvadratnih napak dajo večjo težo večjim odstopanjem, kar ni slučaj pri metrikah absolutnih napak. Kvadratni koren napake ohranja oceno napake v istem velikostnem razredu kot napovedane vrednosti. Metrike relativnih napak kompenzirajo oceno napake z napako, ki je posledica visoke variabilnosti dejanskih vrednosti in jo težko pripišemo neučinkovitosti same metode. Na srečo se v praksi ponavadi izkaže, da je najboljša metoda vedno najbolje ocenjena, ne glede na izbrano metriko.

2.4 Načini primerjave klasifikacijskih in regresijskih metod

Pri iskanju rešitve za problem na področju strojnega učenja se pogosto srečamo z vprašanjem, katera izmed metod je v dani situaciji najbolj ustrezna. Odgovor se zdi preprost: na podlagi prečnega preverjanja ocenimo napake posameznih metod in izberemo tisto, ki se na testnem poskusu izkaže najbolje. Pogosto tak pristop zadostuje, vendar če želimo dokazati, da je za reševanje konkretnega problema izbrana metoda res najboljša, si moramo pomagati s statističnimi testi. V tem poglavju so predstavljeni testi, ki smo jih uporabili pri preverjanju naših hipotez, kakor tudi statistične osnove, na katerih slonijo.

Preden se lotimo opisa statističnih testov, ki se uporabljajo za primerjavo metod na podlagi njihovih ocen, moramo poznati nekaj osnovnih pojmov s področja statistike. Za lažje razumevanje si predstavljajmo, da smo soočeni s klasifikacijskim problemom, kjer nas zanima natančnost klasifikacijske metode. Kot smo omenili že v začetku poglavja, lahko izračunamo natančnost metode kot razmerje med številom uspešno klasificiranih in številom vseh primerov: $f = \frac{n_u}{N}$. Tako dobimo oceno metode za konkreten primer, medtem ko želimo dobiti oceno dejanske natančnosti za poljubno podatkovno množico iz domene zadanega problema. Če si predstavljamo klasifikacijo posameznih primerov kot zaporedje neodvisnih dogodkov, katerih rezultat je bodisi uspešna bodisi neuspešna klasifikacija, opazimo, da imamo opravka z Bernoullijevim procesom. Zanj je značilno, da je prava verjetnost uspešnega izida vedno ista verjetnost p, ne glede na število dogodkov v zaporedju. Torej, če smo na testnem primeru izmerili natančnost f opazovane metode, želimo na podlagi te ocene izračunati pravo natančnost p. Če upoštevamo, da sta povprečna vrednost in varianca posameznega Bernoullijevega poskusa enaka p oziroma $p(1\text{-}p)$ ter da je f naključna spremenljivka Bernoullijevega procesa z N poskusi, potem je njena povprečna vrednost vedno enaka p, medtem ko se varianca zmanjša N-krat na $p(1\text{-}p)/N$. Za velike vrednosti N se porazdelitev f približa normalni (Gaussovi) porazdelitvi.

Za naključno spremenljivko X s povprečno vrednostjo 0 lahko z določeno gotovostjo trdimo, da se nahaja znotraj intervala $[-z, z]$:

$$Pr[-z \leq X \leq z] = c. \tag{2.8}$$

Če imamo opravka z normalno porazdelitvijo naključne spremenljivke X, lahko vrednosti c in pripadajočih intervalov z najdemo v statističnih tabelah, kjer pa so ponavadi podane v nekoliko

spremenjeni obliki, ki izraža gotovost, da se X nahaja izven danega intervala, oziroma samo njegove zgornje meje:

$$Pr[X \geq z] = c. \tag{2.9}$$

Tako podani verjetnosti pravimo tudi enostranska verjetnost, ki pa je zaradi simetričnosti normalne porazdelitve enaka tudi za spodnjo polovico intervala [-z, z]. Vrednosti z so podane v enotah standardnega odklona od povprečja naključne spremenljivke. Če upoštevamo tudi, da mora biti povprečna vrednost naključne spremenljivke enaka 0, lahko v formulo 2.8 vstavimo naključno spremenljivko f, od katere odštejemo p in normaliziramo njeno varianco, tako da jo delimo s standardnim odklonom $\sqrt{p(1-p)/N}$:

$$Pr\left[-z \leq \frac{f-p}{\sqrt{p(1-p)/N}} \leq z\right] = c. \tag{2.10}$$

Če želimo v formuli 2.10 uporabiti vrednost z iz statistične tabele, ki vsebuje enostranske verjetnosti, moramo uporabiti pravilno izbrano verjetnost c (verjetnost, ki ponazarja gotovost, da se naključna spremenljivka nahaja znotraj intervala [-z, z] odštejemo od 1 in dobljeno vrednost prepolovimo). Za izračun zgornje in spodnje meje intervala zaupanja prave verjetnosti naključne spremenljivke f izrazimo p iz formule 2.10 in dobimo:

$$p = \frac{\left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} + \frac{f^2}{N} + \frac{z^2}{4N^2}}\right)}{(1 + \frac{z^2}{N})}. \tag{2.11}$$

Za primerjavo natančnosti dveh različnih metod bi lahko za obe uporabili opisani postopek na isti podatkovni množici in primerjali dobljene verjetnosti. Tako dobljena ocena bi veljala za konkretno podatkovno množico, vendar ne bi nujno predstavljala prave, splošne ocene metod. Za splošno oceno, neodvisno od izbrane podatkovne množice, moramo izračunati več različnih ocen na različnih množicah in primerjati povprečji za obe metodi.

2.4.1 Studentov t-test parnih vzorcev

Imejmo zaporedje ocen učinkovitosti x_1, x_2, . . . , x_k, ki smo jih dobili z zaporednimi ovrednotenji ene metode na različnih podatkovnih množicah iste velikosti in zaporedje ocen y_1, y_2, . . . , y_k, ki smo jih dobili z ovrednotenjem druge metode na paroma istih množicah. Naj bo $\bar{x}$ povprečna vrednost ocen iz prvega zaporedja in $\bar{y}$ povprečna vrednost ocen iz drugega zaporedja. Želimo ugotoviti, ali se $\bar{x}$ signifikantno razlikuje od $\bar{y}$. V primeru ko imamo na voljo dovolj ocen, ima njihova povprečna vrednost ($\bar{x}$) normalno porazdelitev, ne glede na porazdelitev samih ocen. Označimo z μ pravo vrednost povprečja. Če bi poznali varianco σ_x^2 te normalne porazdelitve, tako da bi jo lahko normalizirali, bi lahko dobili meje intervala zaupanja za μ, kot smo opisali v prejšnjem razdelku. Vendar variance ne poznamo, zato jo moramo oceniti na podlagi danih ocen. To storimo tako, da varianco, izračunano iz ocen x_1, x_2, . . ., x_k, delimo s k. Sedaj lahko normaliziramo porazdelitev naključne spremenljivke $\bar{x}$ in njeno povprečno vrednost spremenimo, tako da je enaka 0:

$$X = \frac{\bar{x} - \mu}{\sqrt{\sigma_x^2 / k}} \tag{2.12}$$

Ker smo uporabili oceno variance, za X ne moremo več predvidevati, da ima normalno porazdelitev. Zato za določanje intervala zaupanja ne moremo uporabiti statističnih tabel za normalno porazdelitev, temveč se moramo poslužiti tabel za Studentovo porazdelitev pri stopnji prostosti (*degree of freedom*) k-1. Intervali zaupanja pri Studentovi porazdelitvi so nekoliko širši kot tisti pri normalni porazdelitvi, kar odraža dodatno negotovost, ki je posledica ocene variance. Pri stopnjah prostosti nad 100 so intervali zaupanja že zelo podobni tistim pri normalni porazdelitvi.

Če sedaj želimo primerjati povprečni oceni $\bar{x}$ in $\bar{y}$, najprej izračunamo razlike d_i med paroma pripadajočimi si ocenami: $d_i = x_i - y_i$. Povprečna vrednost teh razlik je enaka razlik obeh povprečnih ocen $\bar{d} = \bar{x} - \bar{y}$ in ima tudi Studentovo porazdelitev s stopnjo prostosti k-1. V primeru ko sta povprečni oceni enaki, je njuna razlika enaka 0 (to je ničelna hipoteza). Če želimo ugotoviti, ali se metodi signifikantno razlikujeta, moramo preveriti, ali se normalizirana razlika njunih povprečnih ocen (imenovana tudi t-statistika) za dano stopnjo zaupanja nahaja izven intervala zaupanja:

$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2/k}}, \tag{2.13}$$

kjer je σ_d^2 varianca razlik ocen.

Kadar ne moremo primerjati metod na paroma enakih množicah, lahko uporabimo splošnejšo različico t-testa, pri čemer namesto povprečne vrednosti razlik ocen $\bar{d}$ uporabimo razliko povprečnih ocen $\bar{x} - \bar{y}$ in ocenimo varianco razlik z $\frac{\sigma_x^2}{k} + \frac{\sigma_y^2}{l}$, kjer sta k in l števili množic, na katerih smo dobili ocene učinkovitosti za prvo oziroma drugo metodo. Pri določanju stopnje prostosti izberemo manjšo izmed vrednosti k in l.

2.4.2 Wilcoxonov test predznačenih rangov

Studentov t-test predvideva, da sta povprečni oceni primerjanih metod $\bar{x}$ in $\bar{y}$ normalno porazdeljeni, česar pa pogosto ne moremo zagotoviti, še posebej če imamo na voljo omejeno število primerov. Takrat si lahko pomagamo z Wilcoxonovim testom predznačenih rangov (znanim tudi kot Wilcoxonov test enakovrednih parov), ki ne zahteva normalne porazdelitve primerjanih vrednosti in predstavlja neparametrično alternativo Studentovemu t-testu.

Začnimo s podobnimi predpostavkami kot pri t-testu. Naj bo $x_1, x_2, \ldots, x_k$ zaporedje ocen učinkovitosti, ki smo jih dobili z zaporednimi ovrednotenji ene metode na različnih podatkovnih množicah in $y_1, y_2, \ldots, y_k$, zaporedje ocen, ki smo jih dobili z ovrednotenjem druge metode na paroma istih množicah. Potem lahko izračunamo razlike med posameznimi pari vrednosti in jih označimo z d_i, $i=1, \ldots, k$. Razlike, ki so enake 0, zavržemo, saj k testu ne pripomorejo. Preostale razvrstimo po velikosti glede na njihove absolutne vrednosti od najmanjše naprej in jim dodelimo range. Rang, ki se dodeli posamezni razliki d_i $(i=1, \ldots, n;\ n \leq k)$ je naravno število, enako zaporedni številki vrednosti d_i v razvrščenem zaporedju. Če je v zaporedju več enakih razlik, se vsem dodeli rang, ki je povprečna vrednost vsote njihovih zaporednih številk (racionalno število). Nato se izračunata vsoti rangov R^+ in R^- vseh razlik, ki so pozitivne, oziroma negativne, ter se na podlagi njiju določi Wilcoxonova statistika, ki je kar enaka manjši izmed obeh vsot:

$$W = min(R^+, R^-). \tag{2.14}$$

Podobno kot smo to storili pri t-testu, moramo za izračun vrednosti z od statistike W odšteti njeno pričakovano povprečno vrednost in jo deliti s standardnim odklonom porazdelitve. Pričakovana povprečna vrednost je enaka polovični vsoti vseh rangov:

$$\mu = \frac{n(n+1)}{4}, \tag{2.15}$$

standardni odklon pa je enak

$$\sigma_w^2 = \sqrt{\frac{n(n+1)(2n+1)}{24}}, \tag{2.16}$$

tako da lahko zapišemo:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}. \tag{2.17}$$

Za preverjanje ničelne hipoteze, da se primerjani metodi v svoji natančnosti bistveno ne razlikujeta, lahko od izbrani stopnji zaupanja poiščemo pripadajočo vrednost z v statističnih tabelah, pri čemer lahko za $n > 25$ (po nekaterih virih $n > 20$ ali celo samo $n > 10$) predvidevamo, da je W normalno porazdeljena.

2.4.3 Friedmanov test

Kadar želimo med seboj hkrati primerjati več metod, pri čemer ne moremo zagotoviti normalne porazdelitve ocen učinkovitosti, lahko uporabimo Friedmanov test, ki je neparametrična alternativa analizi variance (*ANOVA*) za odvisne vzorce. Podobno kot Wilcoxonov test tudi Frriedmanov rangira metode glede na rezultat (oceno napake), pridobljen na več primerih. Če primerjamo m metod na n primerih, test najprej rangira metode za vsak primer posebej, tako da vsakemu rezultatu metode j ($j = 1, \ldots, m$) na primeru i ($i = 1, \ldots, n$) dodeli rang r_{ij}, tako da je vsota vseh rangov enaka:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} = \frac{nm(m+1)}{2}, \tag{2.18}$$

pri čemer se enakim rezultatom na posameznem primeru dodeli povprečna vrednost pozicij, ki jih zasedajo. Friedmanov test nato za vsako metodo sešteje range njenih rezultatov:

$$R_j = \sum_{i=1}^{n} r_{ij} \,, \tag{2.19}$$

ki jih uporabi za izračun statistike Q:

$$Q = \frac{12}{nm(m+1)} \sum_{j=1}^{m} R_j^2 - 3n(m-1). \tag{2.20}$$

Za dovolj veliko število metod (npr. m > 4) ali dovolj veliko število primerov (npr. n > 15) je porazdelitev statistike Q dober približek porazdelitve χ^2(hi-kvadrat) s stopnjo prostosti m-1, tako da lahko preverimo ničelno hipotezo, da so primerjane metode med seboj enakovredne, s pogojem:

$$Pr[\chi^2_{m-1} \geq Q]. \tag{2.21}$$

Kadar preverjamo zelo majhno število metod in/ali je število primerov, na katerih jih primerjamo, majhno, je aproksimacija s porazdelitvijo χ^2 slaba. Takrat je za izračunano statistiko Q potrebno vrednost p poiskati v tabelah za Friedmanov test. Če ničelno hipotezo zavržemo, je priporočljivo izvesti dodatne teste za medsebojne primerjave posameznih metod (npr. Wilcoxonov test).

2.5 Klasifikacijske in regresijske metode

Podatkovne množice se med seboj ne razlikujejo samo po številu atributov, njihovih tipih in številu vzorcev. Pogosto lahko v njih odkrijemo preproste relacijske strukture, kar lahko naknadno izkoristimo za izbiro najbolj primerne klasifikacijske oz. regresijske metode. V določeni podatkovni množici lahko celotno breme klasifikacije leži že enem samem atributu, medtem ko so vsi drugi odveč in ne pripomorejo h kakovosti odločitve. Podobno lahko pri regresiji napovedana vrednost korelira z enim samim atributom iz učne množice. V nekem drugem primeru lahko vsi atributi neodvisno in enakovredno pripomorejo h končnemu rezultatu. Tretja podatkovna množica lahko vsebuje preprosto logično strukturo, ki se jo da učinkovito predstaviti s pomočjo odločitvenega drevesa. Značilnosti četrte množice se morebiti dajo predstaviti z nekaj neodvisnimi pravili, ki določajo vrednost odločitvenega atributa. V peti podatkovni množici bi bilo možno ugotoviti odvisnosti med različnimi podmnožicami

atributov, medtem ko bi kak numerični atribut šeste množice bil linearno odvisen od drugih numeričnih atributov. Takšnih možnosti je veliko, število vseh možnih podatkovnih množic pa je praktično neskončno. Tudi najbolj zmogljiva orodja za podatkovno rudarjenje lahko spregledajo obstoj struktur in odvisnosti znotraj podatkov, če ne uporabijo pravilne metode.

V poglavju 2.2 smo našteli nekatere izmed najbolj uveljavljenih metod nadzorovanega strojnega učenja, ki se uporabljajo na podatkovnih množicah različnih tipov in z različnimi nameni. Nekatere izmed njih so posebej primerne za delo s kategoričnimi atributi, nekatere operirajo samo z numeričnimi, medtem ko se nekatere lahko uporabijo tako za klasifikacijo kot napovedovanje numeričnih vrednosti. V nadaljevanju bomo predstavili osnovne regresijske metode, ki se uporabljajo za numerično predikcijo in smo jih tudi sami uporabili za nadomeščanje manjkajočih vrednosti v podatkovnih zbirkah.

2.5.1 Metoda k-najbližjih sosedov

Učenje po metodi k-najbližjih sosedov (*k-nearest neighbors*) sodi med ti. »lena učenja« (*lazy learning*), saj učnega postopka sploh ni. Celotno delo se opravi med samim postopkom klasifikacije ali regresije. Zaradi tega mora biti učna množica na voljo pri obravnavi vsakega novega vzorca. Predpostavimo, da želimo s pomočjo metode k-NN določiti razred r_x novemu vzorcu u_x. Med učnimi vzorci s pomočjo izbrane metrike poiščemo k najbližjih primerov u_1, ..., u_k in napovemo večinski razred, tj. razred, ki mu pripada največ izmed izbranih sosedov:

$$r_x = arg\,max_{r\in\{v_1,\ldots,v_n\}} \sum_{i=1}^{k} \delta\left(r, r^{(i)}\right), \tag{2.22}$$

kjer je

$$\delta(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}.$$

Pri regresiji dobimo numerično napoved kot povprečno vrednost razreda vseh k najbližjih sosedov:

$$r_x = \frac{1}{k}\sum_{i=1}^{k} r^{(i)}. \tag{2.23}$$

Za parameter k ponavadi izberemo liho število (npr. 1, 3, 5, 7, 15 in 31 so pogoste nastavitve). Če bi bila podatkovna množica dovolj raznovrstna in ne bi vsebovala šuma, potem bi se algoritem najbolje obnesel pri k = 1. S povečanjem parametra k povprečimo napovedi več bližnjih vzorcev in s tem zmanjšamo vpliv morebitnih napačnih napovedi. Po drugi strani pa s tem povečujemo možnost, da svoj delež k rezultatu prispevajo takšni učni primeri, ki se od vzorca, podanega na vhodu, precej razlikujejo. Optimalno vrednost parametra k je zato potrebno eksperimentalno določiti za vsak problem posebej.

Parameter k določa število primerov iz učne množice, ki prispevajo h končni napovedi, in ne velikosti okolice vzorca, za katerega se napoved računa. Tako se okolica dinamično spreminja v odvisnosti od gostote učnih vzorcev. V primeru fiksno izbrane okolice bi v določenih primerih lahko dobili preveč sosednjih učnih vzorcev, medtem ko bi jih v nekaterih drugih situacijah bilo premalo ali celo nič, ki bi bili dovolj blizu.

Pri računanju razdalje med novim vzorcem in učnimi primeri se ponavadi uporablja kar evklidska razdalja. Zvezni numerični atributi se normalizirajo na interval [0, 1], razdalja med dvema vrednostma istega atributa se izračuna kot njuna absolutna razlika. Za diskretne atribute je razdalja med različnima vrednostma enaka 1, med enakima pa 0. Razdaljo med primeroma u_a in u_b, ki vsebujeta *m* atributov, izračunamo z:

$$D(u_a, u_b) = \sqrt{\sum_{i=1}^{m} d(v^{(i,a)}, v^{(i,b)})^2}, \tag{2.24}$$

kjer je za zvezni atribut A_i

$$d\left(v^{(i,a)}, v^{(i,b)}\right) = \left|v^{(i,a)} - v^{(i,b)}\right|$$

ter za diskretni

$$d\left(v^{(i,a)}, v^{(i,b)}\right) = \begin{cases} 0, & v^{(i,a)} \neq v^{(i,b)} \\ 1, & v^{(i,a)} \neq v^{(i,b)} \end{cases}.$$

Seveda je namesto evklidske razdalje velikokrat smiselno uporabiti kakšno drugo metriko, kar pride posebej v poštev v primeru, ko imamo opravka z vrstnimi atributi, oziroma ko lahko na kakšen drugačen način bolje ocenimo sosednost dveh vzorcev. Kadar so nekateri atributi bolj pomembni od

drugih, jih lahko dodatno obtežimo. V ta namen vse atribute najprej ocenimo z ustrezno izbrano mero za ocenjevanje pomembnosti q. Nato izračunamo razdaljo po formuli:

$$D(u_a, u_b) = \sqrt{\sum_{i=1}^{m} q(A_i) d(v^{(i,a)}, v^{(i,b)})^2}. \quad (2.25)$$

Dodatno lahko obtežimo tudi vplive posameznih učnih primerov na napoved, pri čemer je obtežitev odvisna od razdalje učnega primera od vzorca na vhodu. Pogosta izbira je kvadratični vpliv razdalje, pri katerem se razred klasifikacije izračuna z:

$$r_x = arg\, max_{r \in \{v_1, \ldots, v_n\}} \sum_{i=1}^{k} \frac{\delta(r, r^{(i)})}{D(u_x, u_i)^2}, \quad (2.26)$$

napoved pri regresiji pa z:

$$r_x = \frac{\sum_{i=1}^{k} [r^{(i)} / D(u_x, u_i)^2]}{\sum_{i=1}^{k} [1 / D(u_x, u_i)^2]}. \quad (2.27)$$

Če želimo še povečati vpliv razdalje, lahko izberemo npr. eksponentno obtežitev, če želimo vpliv zmanjšati pa linearno obtežitev. V primeru uporabe obteženih učnih vzorcev ni potrebno omejevati števila najbližjih sosedov na k in pri končni napovedi upoštevamo vpliv celotne učne množice. Vpliv zelo oddaljenih vzorcev je namreč zanemarljiv v primerjavi z najbližjimi sosedi.

Zaradi pristopa »lenega učenja« je metoda najbližjih sosedov med fazo klasifikacije računsko in pomnilniško veliko bolj zahtevna kot metode, ki najprej opravijo dejansko učenje na učni množici. Potrebno je obravnavati prav vse vzorce in za vsakega izračunati razdaljo.

2.5.2 Linearna regresija

Linearna regresija je ena izmed temeljnih metod s področja statistike. Predstavlja osnovo mnogih kompleksnih algoritmov in že sama zelo učinkovito rešuje probleme, pri katerih so podatki v podatkovnih množicah vsaj približno linearno odvisni. Cilj linearne regresije je določiti funkcijo, ki

preslika vrednosti zveznih atributov v realno število, zato predstavlja naravno odločitev pri izbiri metode za napovedovanje numeričnih vrednosti. Predstavili bomo različico linearne regresije, ki pri iskanju funkcije uporablja metodo najmanjših kvadratov (*least-squares regression*).

Predpostavimo, da imamo učno množico z *n* učnimi primeri. Vsak učni primer naj vsebuje *a* atributov. Naj bo $\mathbf{v}^T = \left[1\ v^{(1)}\ \cdots\ v^{(a)}\right]$ vektor vrednosti vseh atributov $A_1,\ldots,A_a$, razširjen z elementom 1. Funkcija, ki jo linearna regresija poskuša določiti, je linearna kombinacija vrednosti vseh (ali ustrezno izbranih) atributov:

$$y = f\left(v^{(1)}, \ldots, v^{(a)}\right),$$

$$y = w_0 + \sum_{i=1}^{a} w_1 v^{(i)} = \mathbf{w}^T\mathbf{v}. \tag{2.28}$$

Označimo z $\mathbf{V}$ matriko vseh učnih primerov, razširjenih z elementom 1, ter z **r** vektor razredov (vrednosti odločitvenih atributov) vseh učnih primerov:

$$\mathbf{V} = \begin{bmatrix} 1 & v^{(1,1)} & \ldots & v^{(a,1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & v^{(1,n)} & \ldots & v^{(a,n)} \end{bmatrix}, \qquad \mathbf{r}^T = \left[r^{(1)}\ r^{(2)}\ \cdots\ r^{(n)}\right].$$

Če za vsak učni primer izračunamo kvadrat napake, ki jo na učnem primeru naredi linearna regresija, in nato te napake seštejemo, dobimo metriko vsote kvadratov napak (*sum of squared errors*):

$$SSE = \sum_{j=1}^{n}\left(r^{(j)} - y^{(j)}\right)^2,$$

oziroma če upoštevamo enačbo 2.14:

$$SSE = \sum_{j=1}^{n}\left(r^{(j)} - w_0 - \sum_{i=1}^{a} w^i\, v^{(i,j)}\right)^2. \tag{2.29}$$

Naloga linearne regresije je, da minimizira vrednost metrike *SSE*. Če je izpolnjen pogoj $n > a + 1$, torej če je neodvisnih primerov v učni množici več kot je atributov v posameznem primeru, dobimo minimalno SSE takrat, ko velja: $\mathbf{w} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}\mathbf{r}$ (Rencher, 1995). Rezultat (vektor **w**) je vektor parametrov linearne funkcije, ki jo metoda uporablja za napovedovanje numeričnih vrednosti, pri čemer

se kljub relativni preprostosti lahko odlično izkaže. Ena glavnih pomanjkljivosti linearne regresije po metodi najmanjših kvadratov je njena občutljivost na osamelce oz. odstopajoče vrednosti (*outliers*). Če se temu želimo izogniti, se lahko poslužimo precej bolj robustne različice, ki napako računa po metodi najmanjših kvadratov srednjih vrednosti (*least median squares*) (Rousseeuw, 1984). Seveda pa ne moremo pričakovati, da se bo v primerih, ko so podatki v učni množici nelinearno odvisni, linearna regresija odrezala bolje kot nelinearni regresorji (npr. nevronske mreže).

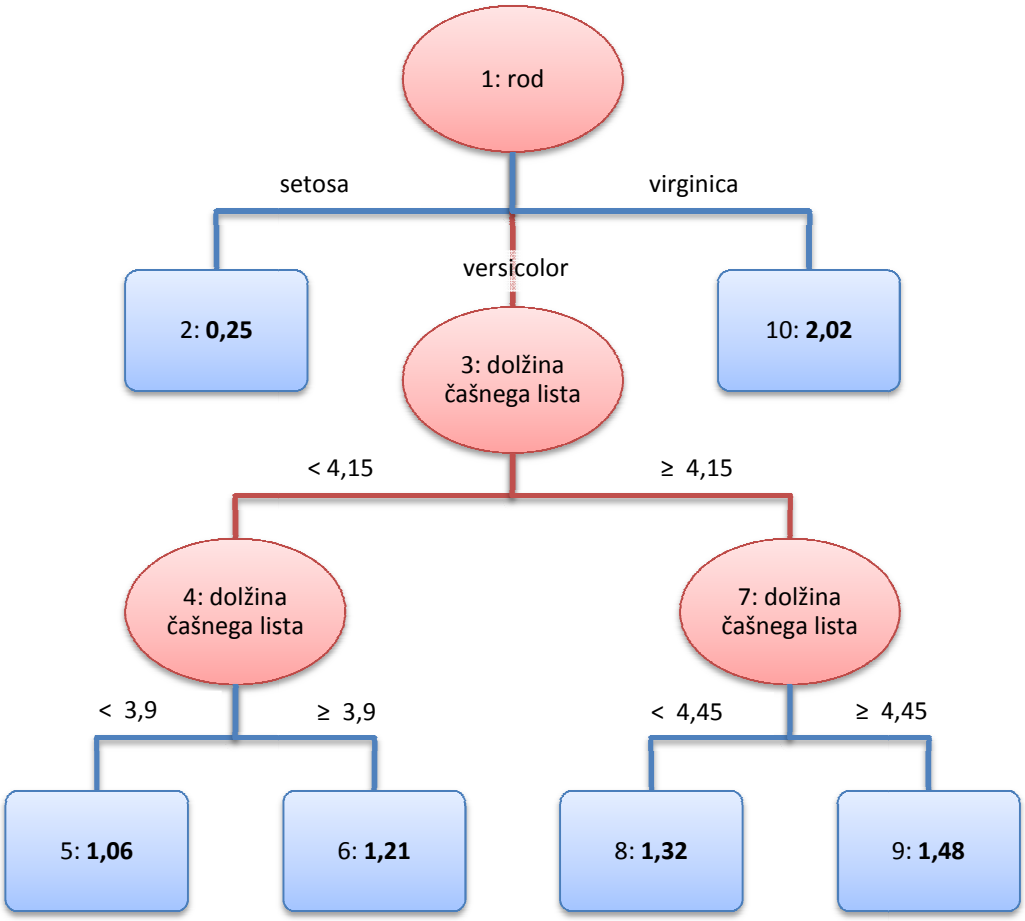
2.5.3 Regresijska drevesa

Regresijska drevesa, katerih naloga je napovedovanje numeričnih vrednosti, se od navadnih odločitvenih dreves razlikujejo v tem, da so v listih drevesa numerične vrednosti, ki predstavljajo povprečno vrednost primerov, ki pridejo do določenega lista, ali pa linearni regresijski modeli, ki napovejo vrednost razrednega atributa. Pri gradnji dreves se uporabljajo učni primeri z zveznim razredom, medtem ko so preostali atributi lahko zvezni ali diskretni.

Regresijsko drevo je sestavljeno iz notranjih vozlišč, ki ustrezajo atributom podatkovne množice, vej, ki ustrezajo podmnožicam vrednosti atributov, in listov, ki ustrezajo funkcijam, ki preslikajo vektor vrednosti atributov v realno število (Slika 2.5-1). Kot smo že omenili v uvodu, je najpreprostejša funkcija v listih kar konstantna vrednost, pogosto pa se uporablja linearna funkcija na podmnožici numeričnih atributov (predhodno se morebitni kategorični atributi pretvorijo v binarne spremenljivke, ki jih algoritem obravnava kot numerične). Takšno regresijsko drevo imenujemo modelno drevo (*model tree*).

Vsaka pot od korena drevesa k listu predstavlja eno pravilo, pri čemer so pogoji v vozliščih konjunktivno povezani. Namen drevesa je, da s temi pravili optimalno razdeli problemski prostor. Gradnja drevesa ponavadi poteka po principu »od zgoraj navzdol« (*top-down*), ki ga je že leta 1986 opisal Quinlan in ga v poenostavljeni obliki sestavljajo trije rekurzivni koraki (Quinlan, 1986):

1. Poišči najboljši atribut, ki v trenutnem vozlišču najbolj učinkovito razdeli problemski prostor glede na učno množico.
2. Razdeli prostor v neprekrivajoče se podmnožice.
3. Za vsakega izmed podprostorov preveri, ali je dosežen ustavitveni pogoj. Če je, iz trenutnega vozlišča ustvari list, drugače rekurzivno ponovi 1. korak.



Slika 2.5-1: Primer regresijskega drevesa za napovedovanje širine venčnega lista perunik na podlagi 2 atributov: roda perunike in dolžine čašnega lista

Izbira najboljšega atributa je ključnega pomena med gradnjo drevesa. Zanesljivost ocene kakovosti atributa je odvisna od števila učnih primerov, zaradi česa ni zaželeno, da se učna množica prehitro razdeli na majhne podmnožice. Zaradi tega algoritmi za gradnjo regresijskih dreves pogosto gradijo binarna drevesa, pri čemer pred izbiro najboljšega atributa izvedejo binarizacijo atributov. Pri zveznih atributih se poišče tista meja znotraj intervala vseh možnih vrednosti danega atributa, ki maksimizira pomembnost atributa.

Za ocenjevanje atributov se najpogosteje uporablja metrika razlike varianc (Breiman, Friedman, Olshen, & Stone, 1984). Varianca zveznega razreda je definirana kot povprečni kvadrat napake:

$$s^2 = \frac{1}{n}\sum_{k=1}^{n}\left(r^{(k)} - \bar{r}\right)^2, \tag{2.30}$$

kjer je $\bar{r}$ povprečna vrednost zveznega razreda z n učnimi primeri:

$$\bar{r} = \frac{1}{n}\sum_{k=1}^{n} r^{(k)}.$$

Oceno pomembnosti atributa A_i dobimo s pomočjo nenegativne pričakovane razlike variance:

$$ds^2(A_i) = \frac{1}{n}\sum_{k=1}^{n}\left(r^{(k)} - \bar{r}\right)^2 - \sum_{j=1}^{n_i}\left(p_{.j}\frac{1}{n_{.j}}\sum_{k=1}^{n_{.j}}\left(r_j^{(k)} - \bar{r}_j\right)^2\right), \tag{2.31}$$

kjer je $n_{.j}$ število učnih primerov z j-to vrednostjo atributa A_i, $p_{.j}$ aproksimacija verjetnosti $p_{.j}=n_{.j}/n$, $r_j^{(k)}$ vrednost zveznega razreda k-tega primera, ki ima j-to vrednost atributa A_i ter $\bar{r}_j$ povprečna vrednost zveznega razreda primerov z j-to vrednostjo atributa A_i:

$$\bar{r}_j = \frac{1}{n_{.j}}\sum_{k=1}^{n_{.j}} r_j^{(k)}.$$

Pri gradnji drevesa se posebna pozornost posveča nižjim nivojem drevesa, kjer vozliščem ustreza majhno število učnih primerov. Ustavitveni pogoj poskuša ustaviti gradnjo, ko le-ta postane nepotrebna ali nezanesljiva. Ker je nezanesljivost težko oceniti med samo gradnjo, se drevo pogosto naknadno oklesti (*postpruning*). Pri postopku naknadnega klestenja ločimo dva različna pristopa: zamenjava poddrevesa (*subtree replacement*) in dvig poddrevesa (*subtree raising*). Drugi pristop je kompleksnejši, časovno zahtevnejši in ne nujno boljši (Witten & Frank, 2005). Pri njem se celotno poddrevo prestavi v višje vozlišče, pri čemer se primeri iz sestrskih vozlišč tega vozlišča razvrstijo po listih novega (manjšega) poddrevesa. Postopek je v svoji vplivni metodi gradnje odločitvenih dreves *C4.5* opisal Quinlan (Quinlan, 1992). Zamenjava poddrevesa je splošnejši postopek, kjer lahko osnovni algoritem opišemo z naslednjimi koraki:

Za vsa notranja vozlišča od spodaj navzgor ponavljaj:

1. Oceni povprečno pričakovano napako regresije v poddrevesih.
2. Oceni pričakovano napako regresije v trenutnem vozlišču.

3. Če je povprečna pričakovana napaka poddreves večja od pričakovane napake vozlišča, potem oklesti poddrevesa in spremeni vozlišče v list.

S pomočjo klestenja se lahko izognemo preveliki prilagoditvi drevesa na primere iz učne množice (*overfitting*), kar privede do večje robustnosti in natančnosti metode. Če algoritem gradi modelno drevo, potem za vsako vozlišče izračuna linearni model:

$$u_0 + u_1 a_1 + u_2 a_2 + \ldots + u_k a_k.$$

Za izračun uteži modela se uporabi navadna linearna regresija na primerih, ki do vozlišč pridejo. Rezultat posameznega pravila, ki se zaključi v določenem listu modelnega drevesa, se naknadno še filtrira skozi postopek glajenja (*smoothing*), tako da se groba ocena linearne funkcije v listu prenaša v višja vozlišča vse do korena, pri čemer se v vsakem vozlišču zgladi po pravilu:

$$p' = \frac{np + kq}{n + k}, \tag{2.32}$$

kjer je p' zglajena napovedana vrednost, ki se prenese v višje vozlišče, p je napovedana vrednost, prenesena v to vozlišče iz nižjega, q je vrednost, ki jo napove model v trenutnem vozlišču, n je število učnih primerov, ki prispejo v trenutno vozlišče in k je konstanta glajenja. Z glajenjem se kompenzira groba nezveznost napovedanih vrednosti med posameznimi listi. Na podlagi poskusov je bilo ugotovljeno, da glajenje občutno poveča natančnost numerične predikcije.

2.6 Ansambli klasifikatorjev

Oculi plus vident quam oculus (več oči vidi več kot eno oko).

Znani rek se glasi: »Več glav več ve«, zato ne čudi, da se je ideja o združevanju posameznih metod strojnega učenja v kompleksnejše sisteme pojavila že pred več desetletji. Dasarathy in Sheela sta že leta 1979 razpravljala o razdelitvi večdimenzionalnega prostora s pomočjo dveh ali več klasifikatorjev (Dasarathy & Sheela, 1979), dobro desetletje kasneje pa sta Hansen in Salamon pokazala, da se splošna klasifikacijska sposobnost nevronske mreže izboljša z uporabo več podobnih mrež, združenih v ansambel (Hansen & Salamon, 1990). Številne raziskave, ki so sledile v preteklih dveh desetletjih, so pripeljale do raznoraznih kombinacij združevanja klasifikatorjev (in regresorjev) v ansambelske sisteme, ki so se v literaturi pojavljali pod raznoraznimi imeni. V nadaljevanju bomo uporabljali izraz »ansambel«, ki se je uveljavil v 90. letih prejšnjega stoletja (Polikar, 2006).

V grobem lahko ansamble po načinu kombiniranja metod razdelimo na 3 skupine:

- uporaba različnih podmnožic učne množice za gradnjo variant iste osnovne metode;
- uporaba različnih parametrov učnega procesa na isti osnovni metodi (npr. različne uteži pri vsaki nevronski mreži v ansamblu);
- uporaba različnih osnovnih metod.

Če bi osnovna metoda strojnega učenja bila popolna, potrebe po združevanju v ansamble ne bi bilo. Vendar obstoj šuma, odstopajočih vrednosti in prekrivajočih porazdelitev podatkov ne dovoljujejo, da bi tak klasifikator bil realen. V najboljšem primeru lahko upamo, da bo uporabljena metoda pravilno opravila svoje delo v večini primerov. Strategija, na kateri slonijo ansambelski sistemi, je, da se ustvari več osnovnih metod, katerih izhod se združi na tak način, da ta kombinacija da boljše rezultate kot vsaka posamezna metoda. Ob tem mora biti izpolnjen pogoj o neodvisnosti posameznih metod, oziroma povedano drugače, posamezne metode se morajo motiti na različnih primerih. Iz tega pogoja namreč sledi posledica, da se s strateškim združevanjem metod lahko zmanjša napaka celotnega sistema. Za metode, ki so edinstvene glede svojih napak, pravimo, da so raznovrstne.

Raznovrstnost ansambelskega sistema lahko dosežemo na različne načine, med katerimi je najbolj pogosto uporabljan pristop uporaba različnih učnih (pod)množic za gradnjo posameznih osnovnih metod. Različne učne množice dobimo z manipulacijo primerkov v osnovni učni množici. Tipični predstavnik tega pristopa je metoda *bagging*, ki jo je leta 1996 predstavil Breiman (Breiman, 1996). Raznovrstnost osnovnih metod je zagotovljena z naključnim kopiranjem učnih vzorcev iz osnovne učne množice, pri čemer kopirani vzorci ostanejo v njej (*bootstrapping*). Posamezen vzorec se lahko v tako ustvarjeni podmnožici večkrat ponovi ali pa se sploh ne pojavi. Rezultati posameznih metod se v enotno napoved združijo s pomočjo večinskega glasovanja.

Breiman je leta 2001 na osnovi metode *bagging* razvil naključne gozdove (*random forests*) (Breiman, 2001). Kot osnovno metodo ansambla je uporabil odločitveno drevo, pri čemer se, kot pri *bagging,* učna množica vsakega drevesa določi z naključnim kopiranjem učnih vzorcev z nadomeščanjem. Pri gradnji posameznih dreves se v vsakem vozlišču naključno izbere vnaprej določeno število atributov iz množice vseh atributov posameznega vzorca, izmed katerih se nato poišče najboljši atribut za razdelitev problemskega prostora v trenutnem vozlišču. Drevesa se ne klestijo, končna odločitev pa je dobljena z večinskim glasovanjem.

Enega najpomembnejših mejnikov pri razvoju ansambelskih sistemov predstavlja metoda *boosting*, ki jo je leta 1990 opisal Schapire in jo utemeljil na ideji, da je možno vsak šibek učni algoritem (katerega natančnost je le malenkost boljša od naključnega odločanja) nadgraditi v močnega (Schapire, 1990). Kot pri *bagging* se tudi pri *boosting* med gradnjo osnovnih metod uporabljajo različne podmnožice učne množice, vendar je izbira vzorcev strateško naravnana. *Boosting* zgradi tri različne šibke metode, pri čemer za gradnjo prve uporabi naključno izbrano podmnožico učnih vzorcev. Za gradnjo druge metode se učni vzorci izberejo tako, da jih prva šibka metoda polovico klasificira pravilno in polovico nepravilno. Tretja metoda se zgradi na vzorcih, na katerih se prvi dve metodi ne strinjata. Končni rezultat ansambla je dobljen na osnovi tristranskega večinskega glasovanja. Schapire je dokazal, da je napaka takšnega ansambla vedno manjša od napake najboljše izmed treh osnovnih metod, če so le napake osnovnih metod manjše od 0,5 (torej manjše kot bi bila napaka pri naključnem dvosmernem odločanju). Z rekurzivnim izvajanjem algoritma je tako možno ustvariti močan klasifikator. Leta 1997 sta Freund in Schapire predstavila izboljšavo osnovnega algoritma *boosting*, imenovano *AdaBoost* (*ADAptive BOOSTing*), ki je odpravila večino pomanjkljivosti osnovne različice in se je uveljavila kot ena izmed najbolj priljubljenih ansambelskih metod (Freund & Schapire, 1997).

V nadaljevanju bomo podrobneje predstavili metodo *bagging*, ki smo jo v doktorski nalogi uporabili kot eno izmed metod za primerjavo z našim rotacijskim regresijskim gozdom.

2.6.1 Bagging

Kot smo že omenili v prejšnjem razdelku, predstavlja raznolikost osnovnih metod osnovo za uspešno delovanje ansambelskega sistema. *Bagging* v ta namen gradi osnovne klasifikatorje (regresorje) na različnih (enako velikih) učnih množicah. Čeprav bi lahko pričakovali, da bodo tako zgrajeni klasifikatorji enako klasificirali večino testnih vzorcev, se v praksi izkaže, da temu ni tako. Ravno nasprotno, če je učna množica relativno majhna in je osnovna metoda nestabilna, so razlike med posameznimi napovedmi osnovnih metod precejšnje ter se rezultati ansambla izkažejo za presenetljivo dobre (Breiman, 1996). Pogosto uporabljani osnovni metodi sta tako zaradi svoje nestabilnosti nevronska mreža in odločitveno (regresijsko) drevo. Gradnja regresijskega drevesa, ki smo jo predstavili v razdelku 2.5, je nestabilen postopek: že majhne razlike v učni množici lahko privedejo do izbire drugega atributa v določenem vozlišču, kar spremeni strukturo poddreves pod tem vozliščem. Isto velja tudi za odločitvena drevesa in sklepamo lahko, da se v različnih drevesih ansambla posamezni testni vzorci lahko klasificirajo v različne razrede.

Rezultat klasifikacije s pomočjo ansambla je razred, ki dobi največ glasov v enakovrednem glasovanju posameznih osnovnih metod (Slika 2.6-1).

gradnja ansambla:

VHOD: osnovni klasifikator $\boldsymbol{K}$, število osnovnih klasifikatorjev $\boldsymbol{N}$, učna množica $\boldsymbol{U}$

ZA $\boldsymbol{k}$ = 1 do $\boldsymbol{N}$:

- iz učne množice $\boldsymbol{U}$ s številom vzorcev $\boldsymbol{n}$ naključno IZBERI Z NADOMEŠČANJEM podmnožico $\overline{\boldsymbol{U}}$ z $\boldsymbol{m}$ vzorci ($\boldsymbol{m} \leq \boldsymbol{n}$)
- na učni množici $\overline{\boldsymbol{U}}$ ZGRADI klasifikator $\boldsymbol{K_k}$ tipa $\boldsymbol{K}$

klasifikacija:

ZA $\boldsymbol{k}$ = 1 do $\boldsymbol{N}$:

- NAPOVEJ razred $\boldsymbol{r_i}$ vhodnega vzorca $\boldsymbol{v}$: $\boldsymbol{r_i} = \boldsymbol{K_k(v)}$

VRNI razred $\boldsymbol{r}$, ki je dobil največ glasov: $max_r\left(\sum_{k=1}^{N} K_k(v) = r\right)$

Slika 2.6-1: Psevdokod algoritma ansambla *bagging*

Učne možice, uporabljene pri gradnji posameznih osnovnih metod, se ustvarijo iz osnovne učne množice z naključno izbiro z nadomeščanjem (*bootstrapping*). To pomeni, da se posamezni vzorec osnovne učne množice lahko večkrat ponovi v tako ustvarjeni podmnožici, ali pa se sploh ne pojavi. Verjetnost, da se pojavi v učni podmnožici z n elementi, je enaka:

$$p = 1 - \left(1 - \frac{1}{n}\right)^n, \tag{2.33}$$

kar pri $n \to \infty$ limitira k $1 - \frac{1}{e}$, oziroma približno 0,632. Tako ustvarjene podmnožice se pogosto občutno prekrivajo in seveda med seboj niso neodvisne. Ne glede na to se izkaže, da je ansambel, sestavljen iz metod, grajenih na takšnih učnih množicah, pogosto signifikantno boljši od osnovne metode, zgrajene na celotni učni množici, ter ni nikoli občutno slabši.

Za nas je zanimivo predvsem dejstvo, da se *bagging* poleg klasifikacije lahko uspešno uporabi tudi za napovedovanje numeričnih vrednosti. Osnovna metoda je v tem primeru lahko regresijsko ali modelno drevo, pri čemer je edina razlika v glasovanju posameznih metod. Namesto večinskega glasovanja, ki pri regresiji ni možno zaradi ponavadi povsem različnih napovedi osnovnih metod, se napovedane numerične vrednosti povprečijo. V teoriji je možno dokazati, da se s povprečenjem napovedi, pridobljenih s pomočjo metod, zgrajenih na neodvisnih učnih množicah, pričakovana povprečna kvadratna napaka napovedi vedno zmanjša (Witten & Frank, 2005).

3 Manjkajoče vrednosti

3.1 Vzroki in posledice

Manjkajoče vrednosti v podatkovnih množicah, pridobljenih s pomočjo eksperimentov ali opazovanj, pogosto predstavljajo neizogiben problem, ne glede na področje empiričnih raziskav. Vzroki njihovega nastanka so številni in so lahko odvisni ali neodvisni od samih podatkov. Nekateri tipični primeri so:

- napake med vnosom (npr. izpuščen datum),
- namerno neodgovorjeno vprašanje v anketi (npr. o starosti),
- drag postopek pridobivanja podatkov (npr. uporaba dragih snovi med izvajanjem eksperimenta),
- izguba podatkov zaradi napake v postopku (npr. prah na mikromrežnem čipu),
- nezmožnost izvedbe vseh meritev (npr. v kliničnem preizkusu udeleženec ni prisoten ob dogovorjenem terminu), itd.

Efron je v definicijo manjkajočih vrednosti vključil tudi latentne (neopazovane) vrednosti (Efron 1994), kar pri statistični analizi lahko omogoči natančnejše modeliranje in izpeljavo popolnejše računske metodologije. Pri naši nalogi se z latentnimi vrednostmi ne bomo srečevali in se bomo pri definiciji manjkajočih vrednosti omejili samo na opazovane manjkajoče vrednosti in metode, ki se z njimi ukvarjajo.

Namen katerekoli analize podatkov je, da pride do določenih sklepov na podlagi podatkov, ki so na voljo. Manjkajoče vrednosti lahko onemogočijo uspešno sklepanje (inferenco), če vplivajo na podatke na način, da ti ne predstavljajo reprezentančnega primera populacije, iz katere so bili povzeti. Drugi, očitnejši razlog za nadomeščanje manjkajočih vrednosti je nujnost popolnih podatkov pri nadaljnji analizi (npr. genskih ekspresij). Pomembno se je tudi zavedati, da ko imamo enkrat opravka z manjkajočimi podatki, bodo ti v vsakem primeru obravnavani pri nadaljnji statistični analizi, zato je smiselno manjkajoče vrednosti še pred tem nadomestiti na najustreznejši možni način. Preprosti postopki, kot je brisanje pomanjkljivih vzorcev in druge *ad-hoc* metode, lahko namreč naredijo več škode kot koristi (Little & Rubin, 1987), (Schafer, 1997). Z grobimi ocenami manjkajočih vrednosti sicer omogočijo analizo nad vsemi vzorci, vendar pogosto privedejo do pristranskega sklepanja, ki ne odraža dejanskega stanja

populacije. Izbira prave metodologije pri analizi manjkajočih podatkov je v največji meri odvisna od podatkov samih, oziroma od vzroka nastanka manjkajočih vrednosti in njihovega deleža. Za razumevanje razlik med tipi manjkajočih vrednosti je potrebno le-te ločiti na podlagi njihovega nastanka.

V nadaljevanju bodo predstavljeni trije različni mehanizmi nastanka manjkajočih vrednosti, kot jih je opisal Rubin (Rubin, 1976).

3.2 Delitev glede na mehanizem nastanka

Mehanizem nastanka manjkajočih vrednosti je proces, pri katerem postanejo podatki nepopolni. Od mehanizma je odvisno, ali je uporaba določene metode za nadomeščanje manjkajočih vrednosti ustrezna. Metoda, ki se odlično obnese pri enem tipu manjkajočih vrednosti, lahko popolnoma odpove pri drugem in privede do napačnih statističnih inferenc. Rubin je že leta 1976 razdelil manjkajoče vrednosti glede mehanizma njihovega nastanka v tri skupine, ki so bile podrobneje opisane desetletje kasneje (Little & Rubin, 1987):

- povsem naključno manjkajoči podatki (*Missing Completely at Random* – MCAR), pri katerih je vzrok nastanka povsem neodvisen od opazovanih in manjkajočih vrednosti;
- naključno manjkajoči podatki (*Missing at Random* – MAR), pri katerih je vzrok nastanka neodvisen od samih manjkajočih vrednosti, vendar je odvisen od preostalih opazovanih podatkov;
- nenaključno manjkajoči podatki (*Not Missing at Random* – NMAR), pri katerih je vzrok nastanka odvisen od manjkajočih vrednosti samih.

Preden se lotimo bolj formalne definicije posameznih mehanizmov, vpeljimo naslednjo notacijo:

- Y – naključna spremenljivka (vektor) vseh vrednosti posameznega vzorca podatkovne množice;
- R – naključna spremenljivka (vektor), ki je indikator manjkajočih vrednosti, pripadajočih Y;
- (y, r) – par dejanskih vektorjev manjkajočih vrednosti in njim pripadajočih indikatorjev, kjer imajo komponente vektorja r vrednost 1, če je pripadajoča komponenta vektorja y znana, in vrednost 0, če je pripadajoča komponenta vektorja y manjkajoča vrednost.

3.2.1 Povsem naključno manjkajoči podatki (MCAR)

Za mehanizem MCAR lahko zapišemo pogojno verjetnost, da je spremenljivka R enaka določenemu vektorju r, kot:

$$P(R = r|y) = P(R = r). \tag{3.1}$$

Torej je pogojna verjetnost pojavitve manjkajočih vrednosti v vzorcu y neodvisna od vrednosti y, oziroma vrednosti komponent y (ali poznavanje le-teh) ne vplivajo na verjetnostno porazdelitev spremenljivke R. Povedano še nekoliko drugače, manjkajoče vrednosti tipa MCAR se pojavljajo neodvisno od drugih vrednosti vzorca kakor tudi od svojih dejanskih vrednosti.

Manjkajoči podatki, nastali po mehanizmu MCAR, pri nadaljnji analizi dovoljujejo uporabo istih statističnih metod, ki bi jih uporabili, če manjkajočih podatkov ne bi bilo. Sicer lahko pričakujemo izgubo informacij, vendar lahko na podlagi popolnih vzorcev pridemo do veljavnih statističnih inferenc (dovolimo si lahko predhodno brisanje nepopolnih vzorcev).

Primeri manjkajočih vrednosti, nastalih po mehanizmu MCAR:

- nenamerno spregledano vprašanje na anketnem listu,
- naključno uničenje biološkega vzorca pred izvajanjem meritve (npr. med transportom),
- prah ali praska na mikromrežnem čipu, itd.

3.2.2 Naključno manjkajoči podatki (MAR)

Pogojno verjetnost, da je spremenljivka R enaka določenemu vektorju r, lahko za mehanizem MAR zapišemo kot:

$$P(R = r|y) = P\big(R = r|y_{op}\big), \tag{3.2}$$

kjer je y_{op} znani del vektorja y.

To pomeni, da je verjetnost pojavitve določenega indikacijskega vzorca manjkajočih vrednosti odvisna od vrednosti znanih komponent vektorja y. Torej se manjkajoče vrednosti tipa MAR pojavljajo odvisno od vrednosti ene ali več znanih komponent pripadajočega vzorca podatkovne množice, ne pa

tudi od svojih dejanskih vrednosti. Vsekakor je ime mehanizma nekoliko zavajajoče, saj dejansko ne moremo govoriti o naključnem nastanku manjkajočih vrednosti.

Pri nadaljnji statistični analizi se lahko uporabljajo metode, ki temeljijo na verjetju (*likelihood*), medtem ko lahko preostale metode privedejo do zavajajočih sklepanj (npr. spremenjena povprečna vrednost). Zaradi tega je pomembno, da se mehanizem MAR pravilno identificira in ga ne zamenjamo za MCAR.

Primeri manjkajočih vrednosti, nastalih po mehanizmu MAR:

- neodgovorjeno vprašanje na anketnem listu, ki se nanaša samo na del populacije (npr. mlajše od 30 let),
- izključitev preiskovancev iz dela kliničnega preizkusa na podlagi določene meritve (npr. raven holesterola in meritev krvnega pritiska),
- pogojna dodatna meritev (npr. če sta prvi dve meritvi zelo različni, se opravi še tretja), itd.

3.2.3 Nenaključno manjkajoči podatki (NMAR)

Kadar ne moremo trditi, da je mehanizem nastanka manjkajočih vrednosti enak MCAR ali MAR, imamo opravka z nenaključno manjkajočimi podatki. V literaturi lahko zasledimo različne kratice, s katerimi so avtorji označili ta najzahtevnejši mehanizem. Poleg NMAR, npr. (Jamshidian & Bentler, 1999), se najpogosteje uporabljata kratici MNAR (*Missing Not at Random*), npr. (Kenward & Molenberghs, 1999) in NI (*Non-Ignorable*), npr. (Little & Rubin, 2002).

Nastanek manjkajočih vrednosti tipa NMAR je odvisen od dejanskih manjkajočih vrednosti. To ima neposreden vpliv na porazdelitev vrednosti spremenljivke, pri kateri manjkajoče vrednosti nastopajo (možna je opazna sprememba povprečne vrednosti in standardnega odklona). Tudi če upoštevamo vse znane opazovane podatke, vzrok manjkanja neznanih vrednosti še vedno leži v njih samih. Zaradi tega je izmed vseh treh mehanizmov najtežje ustvariti statistični model za mehanizem NMAR. Za zagotovitev veljavnega statističnega sklepanja je potrebno v model združiti znane podatke in čim bolj ustrezen model mehanizma nastanka manjkajočih vrednosti.

Na podlagi samih podatkov ne moremo ugotoviti, ali je mehanizem tipa NMAR (sicer lahko razločimo med MCAR in MAR). Pogosto je potrebno opraviti ločene analize, pod predpostavko

mehanizma MAR in NMAR in z uporabo različnih modelov, kar je lahko časovno in finančno zelo zahtevno.

Primeri manjkajočih vrednosti, nastalih po mehanizmu NMAR:

- neodgovorjeno vprašanje na anketnem listu, ki ni odgovorjeno zaradi vrednosti samega odgovora (npr. osebe z višjimi dohodki ne želijo zaupati podatka o dohodku),
- neudeležba preiskovancev v kliničnem preizkusu zaradi zdravstvenega stanja, povezanega s preizkusom (npr. zaradi hude depresije se preiskovanec ne udeleži ocenjevanja stopnje depresije),
- merilni inštrument ne omogoča skrajnih meritev (npr. premer drevesa je večji od razpona merilnih klešč), itd.

3.3 Načini ravnanja z manjkajočimi vrednostmi

Metode, ki se ukvarjajo s problematiko manjkajočih vrednosti, obsegajo širok spekter možnih pristopov, od najbolj preprostih, kot je brisanje nepopolnih vzorcev, do najbolj zapletenih, kot je npr. večkratno vstavljanje, pri katerem mora uporabnik večkrat ponoviti celoten postopek naknadne statistične analize, s pomočjo katere želi priti do začrtanih sklepov. Izbira ustrezne metode je v veliki meri odvisna od tipa podatkov in mehanizma nastanka manjkajočih vrednosti.

Za ugotovitev tipa mehanizma je potrebno pridobiti informacije o manjkajočih podatkih. V ta namen je možno ustvariti in preveriti točno določeno teorijo, ki temelji na nekem predznanju, ali pa pridobiti dodatne informacije, npr. na podlagi naknadnih meritev. S pomočjo Studentovega t-testa je možno preveriti, ali mehanizem nastanka ni MCAR (s primerjanjem porazdelitve vrednosti spremenljivk, razporejenih v dve skupini glede na manjkajoče vrednosti).

Nekatere izmed najpogosteje uporabljanih metod so tudi najbolj preproste in ustvarijo popolno podatkovno množico z vstavljanjem povprečnih vrednosti ali kar z brisanjem nepopolnih vzorcev. Priljubljene so predvsem zaradi preproste implementacije, vendar lahko privedejo do zelo zavajajočih rezultatov (Little & Schenker, 1995). Uporabne so samo na manjkajočih vrednostih tipa MCAR, če ohranjanje standardnega odklona ni potrebno. Naprednejše metode, ki temeljijo na linearni regresiji, iskanju najbližjih sosedov in druge metode enkratnega vstavljanja (*single-imputation methods*) so primerne tudi pri zahtevnejšem mehanizmu MAR, če odstotek manjkajočih vrednosti ni ekstremno visok.

Lahko se uporabijo tudi pri majhnih odstotkih (do 5%) manjkajočih vrednosti tipa NMAR (Scheffer, 2002). Metode, ki so v splošnem edine primerne za nadomeščanje manjkajočih vrednosti, nastalih po mehanizmu NMAR, so metode večkratnega vstavljanja. Njihova glavna pomanjkljivost je njihova zahtevnost.

V nadaljevanju so podrobneje predstavljeni najbolj priljubljeni pristopi, ki smo jih tudi mi uporabili za primerjavo z našo metodo nadomeščanja manjkajočih vrednosti.

3.3.1 Brisanje in nadomeščanje z 0

Brisanje nepopolnih vzorcev (*listwise/case deletion*) je najpogosteje uporabljana rešitev za eliminacijo manjkajočih vrednosti. Če je mehanizem nastanka manjkajočih podatkov MCAR, potem se z brisanjem vzorcev ohranja povprečna vrednost in varianca podatkov, vendar se zmanjša obseg množice, kar ima za posledico zmanjšanje signifikance statističnih testov. Pri mehanizmu MAR brisanje nepopolnih vzorcev v večini primerov privede do pristranskega sklepanja (Graham & Donaldson, 1993) in je seveda neprimerno pri manjkajočih podatkih tipa NMAR.

Nadomeščanje z 0 je metoda, ki se uporablja pri nadomeščanju neznanih numeričnih vrednosti, kadar ne želimo odstraniti nepopolnih vzorcev. Uporabna je le takrat, ko je odstotek manjkajočih vrednosti majhen in bi brisanje pomanjkljivih vzorcev privedlo do pretiranega zmanjšanja obsega množice. Občutno vpliva na povprečno vrednost in varianco podatkov.

3.3.2 Nadomeščanje s povprečno vrednostjo

Metoda nadomeščanja s povprečno vrednostjo nadomesti vse manjkajoče vrednosti določene spremenljivke s povprečno vrednostjo vseh znanih vrednosti te spremenljivke. Tako ohranja povprečno vrednost nespremenjeno, ne glede na odstotek manjkajočih vrednosti. Kadar imamo opravka z asimetrično porazdelitvijo vrednosti spremenljivke, je povprečna vrednost lahko v večini primerov slab nadomestek manjkajoče vrednosti. Še posebej to pride do izraza pri mehanizmih MAR in NMAR. Z večanjem odstotka manjkajočih vrednosti se manjša skupna varianca spremenljivke, saj je varianca vseh nadomeščenih vrednosti enaka 0. Posledica tega je tudi podcenjevanje korelacije z drugimi spremenljivkami. Iz teh razlogov je Graham metodo označil za »nesprejemljivo« (Graham, 2003).

3.3.3 Metode enkratnega vstavljanja (single-impute methods)

V to skupino lahko uvrstimo vse netrivialne metode, ki z nadomeščanjem manjkajočih vrednosti ustvarijo eno popolno podatkovno množico. Ločimo dva glavna pristopa k napovedovanju manjkajočih vrednosti: iskanje podobnih vzorcev in linearno regresijo. Najpogosteje uporabljani metodi, ki jih lahko uvrstimo v prvo skupino, sta metoda k-najbližjih sosedov, ki smo jo že predstavili v 2. poglavju in ti. »hot-deck« vstavljanje, pri katerem se posamezna manjkajoča vrednost nadomesti z ustrezno znano vrednostjo najbolj podobnega vzorca. Obe predstavljata korak naprej pri ohranjanju porazdelitve spremenljivk, kar posledično pomeni tudi manjše zmanjšanje variance.

Tudi linearno regresijo smo že opisali. Poleg osnovne linearne regresije se uporabljajo tudi številne različice metode srednjih kvadratov (Brock, Shaffer, Blakesley, Lotz, & Tseng, 2008). Njihova uspešnost je odvisna predvsem od modela, ki ga zgradijo, torej posledično od učne množice. Večanje odstotka manjkajočih vrednosti tako zmanjšuje zmožnost generalizacije. Nekatere različice zato uporabljajo dodatno stohastično komponento, ki skrbi za večanje variance na podlagi negotovosti napovedi.

V splošnem se metode enkratnega vstavljanja obnesejo bolje kot tradicionalne metode, nimajo večjih težav z nadomeščanjem manjkajočih vrednosti tipa MCAR, so uporabne pri manjših odstotkih manjkajočih vrednosti tipa MAR in le v izjemnih primerih primerljive z metodami večkratnega vstavljanja, kadar je mehanizem nastanka manjkajočih vrednosti NMAR.

3.3.4 Večkratno vstavljanje

Tako kot pri večini metod enkratnega vstavljanja se pri večkratnem vstavljanju manjkajoče vrednosti katerekoli spremenljivke napovejo na podlagi obstoječih vrednosti drugih spremenljivk, oz. atributov. Postopek nadomeščanja manjkajočih vrednosti se ponovi večkrat, kar privede do več različnih popolnih podatkovnih množic z napovedanimi manjkajočimi vrednostmi. Standardna statistična analiza se izvede na vsaki izmed teh množic, kar proizvede več rezultatov analize. Ti rezultati se nato združijo v skupen rezultat splošne analize.

Postopek večkratnega vstavljanja napove manjkajoče vrednosti s ciljem obnove naravne variabilnosti manjkajočih podatkov, pri čemer poskuša v napovedi vključiti tudi negotovosti, ki so posledica ocenjevanja manjkajočih vrednosti. Prvotno variabilnost ohranja z upoštevanjem mehanizma

nastanka manjkajočih vrednosti, medtem ko negotovost napovedi vpelje z ustvarjanjem raznolikih popolnih podatkovnih množic in ocenjevanjem tako nastale variabilnosti.

Namen nadomestnih napovedanih vrednosti pri večkratnem vstavljanju tako ni ugibanje čim bolj natančnih posameznih napovedi, temveč ohranjanje skupne variance znotraj opazovane populacije, pri čemer se obdržijo relacije z drugimi spremenljivkami. Cilj večkratnega vstavljanja je zastavljen drugače kot pri enkratnem vstavljanju, saj se fokus ne nahaja na posameznih napovedih, temveč na ohranjanju pomembnih značilnosti podatkov, kot so npr. povprečne vrednosti, variance in regresijski parametri.

Čeprav je princip analize več hkratnih nadomestnih podatkovnih množic, ki jih zahteva postopek večkratnega vstavljanja, za končnega uporabnika lahko zastrašujoč, so rezultati tega skoraj vedno vredni, če je namen vstavljanja manjkajočih vrednosti izvajanje nadaljnje statistične analize na popolnih podatkih. Kot alternativa je uporaba metod enkratnega vstavljanja smiselna predvsem, kadar so manjkajoče vrednosti tipa MCAR, oziroma uporabnika zanima napoved dejanskih vrednosti posameznih vzorcev.

4 Rotacijski gozd

V 2. poglavju smo omenili naključne gozdove, ki jih je razvil Breiman (Breiman, 2001). Kot osnovno metodo ansambla je uporabil odločitveno drevo, kar je razlog, da je ansambel poimenoval »naključni gozd« (*random forest*). Nadgradil je metodo *bagging*, tako da je pri gradnji posameznih dreves dodal korak, pri katerem se v vsakem vozlišču drevesa iz naključno izbrane podmnožice vseh atributov poišče najboljši atribut za razdelitev problemskega prostora v trenutnem vozlišču. Ta navidez majhna sprememba je povzročila, da se je povečala raznovrstnost osnovnih klasifikatorjev, pri čemer se njihova natančnost ni bistveno poslabšala. Kljub temu so komparativne študije pokazale, da je natančnost ansambelskega sistema, zgrajenega po metodi *AdaBoost*, v povprečju še vedno nekoliko višja od natančnosti, ki jo doseže naključni gozd (Bauer & Kohavi, 1999), (Banfield, Hall, Bowyer, Bhadoria, Kegelmeyer, & Eschrich, 2004). Ena izmed razlag za večjo uspešnost metode *AdaBoost*, posebej pri majhnem številu osnovnih klasifikatorjev, pravi, da vzrok višje natančnosti leži v večji raznolikosti osnovnih klasifikatorjev (Margineantu & Dietterich, 1997). Na podlagi te predpostavke so Rodríguez, Kuncheva in Alonso razvili novo ansambelsko metodo, ki temelji na Breimanovih naključnih gozdovih in hkrati poskuša dodatno povečati raznolikost osnovnih klasifikatorjev. Ansambel so poimenovali rotacijski gozd (Rodríguez, Kuncheva, & Alonso, 2006). Bistvena razlika med naključnimi gozdovi in rotacijskim gozdom leži v koraku transformacije prostora, ki je vključen v metodo rotacijskih gozdov. Transformacija prostora se izvede z uporabo metode glavnih komponent (angl. Principal Component Analysis), ki jo imenujemo tudi PCA (Hotelling, 1933).

4.1 Gradnja rotacijskega gozda

Preden se lotimo opisa postopka gradnje rotacijskega gozda, vpeljimo potrebno notacijo. Naj bo $x = [x_1, \dots, x_n]$ podatkovna točka, opisana z n atributi in naj bo X podatkovna množica, ki vsebuje N učnih vzorcev, predstavljenih v obliki matrike $n \times N$. Vektor $Y = [y_1, \dots, y_N]^T$ naj predstavlja vrednosti odločitvenega atributa, kjer je y_j vrednost iz nabora možnih vrednosti odločitvenega atributa $\{\omega_1, \dots, \omega_c\}$. Z D_1,...,D_L označimo klasifikatorje v ansamblu in z F množico vseh atributov.

Kot pri večini ansamblov moramo število osnovnih klasifikatorjev L določiti že pred gradnjo. Za gradnjo posameznega klasifikatorja D_i izvedemo naslednje korake:

1. Razdelimo F v K naključnih podmnožic (K lahko kot parameter algoritma nastavimo na želeno vrednost). Podmnožice so lahko prekrivajoče, vendar zaradi večje raznolikosti klasifikatorjev uporabimo neprekrivajoče podmnožice atributov. Za enostavnejše razumevanje predpostavimo, da je n večkratnik K in zato vsaka podmnožica vsebuje $M = n/K$ atributov.
2. Z $F_{i,j}$ (j = 1, ..., K) označimo j-to podmnožico atributov za učno množico klasifikatorja D_i. Za vsako podmnožico $F_{i,j}$ naključno izberemo neprazno podmnožico vrednosti odločitvenega atributa, nato pa iz osnovne učne množice X z nadomeščanjem izberemo 75% vzorcev. Sedaj nad temi vzorci z upoštevanjem samo atributov iz pripadajoče podmnožice $F_{i,j}$ izvedemo transformacijo prostora z analizo glavnih komponent (PCA). Koeficiente glavnih komponent $a_{i,j}^1, \dots, a_{i,j}^M$ shranimo v vektorje velikosti $M \times 1$.
3. Tako dobljene vektorje lahko zapišemo v obliki naslednje redke »rotacijske« matrike R_i:

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \dots, a_{i,1}^{(M)} & [0] & \cdots & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)}, \dots, a_{i,2}^{(M_2)} & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \dots, a_{i,K}^{(M_K)} \end{bmatrix}. \quad (4.1)$$

Novo (rotirano) učno množico klasifikatorja D_i dobimo tako, da atribute (stolpce v R_i) razporedimo v enak vrstni red, kot je bil na začetku (v F) in takšno preurejeno matriko označimo z R_i^a. Učno množico za gradnjo klasifikatorja D_i dobimo kot XR_i^a.

Po določitvi učnih množic sledi še postopek gradnje dreves. V fazi klasifikacije klasifikator D_i dodeli vzorcu x verjetnost, da x pripada razredu ω_j v obliki $d_{i,j}(xR_i^a)$. Nato za vsak ω_j (j = 1,...,c) izračunamo:

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(xR_i^a). \quad (4.2)$$

Vrednost klasifikacije (odločitvenega razreda) vzorca x predstavlja ω_j z najvišjo vrednostjo $\mu_j(x)$.

4.1.1 Analiza osnovnih komponent

Metoda glavnih (ali osnovnih) kompnent (*Pricnipal Component Analysis - PCA*) je ena najpogosteje uporabljenih multivariatnih statističnih metod. Zasnoval jo je Karl Pearson (Pearson, 1901), največ zaslug

za nadaljnji razvoj pa ima Hotelling (Hotelling, 1933). Analiza osnovnih komponent nam pomaga razkriti interno strukturo podatkov na način, ki najbolje ponazarja varianco le-teh, tako da izvede redukcijo števila dimenzij v podatkih. S pomočjo linearnih transformacij zagotavlja najbolj optimalno preslikavo iz visoko dimenzionalnega prostora v nižje dimenzionalni prostor.

Recimo, da si želimo iz začetnih m-dimenzionalnih podatkov ustvariti njihovo n-dimenzionalno predstavitev, pri čemer velja $n < m$. Ta cilj *PCA* doseže tako, da poišče n najbolj informativnih vektorjev za dane podatke. Vektorji so lahko poljubna linearna kombinacija posameznih atributov. Najbolj informativni vektorji so tisti, ki zajamejo maksimalno varianco v podatkih. Prva osnovna komponenta (vektor), bo tako tista linearna kombinacija osnovnih atributov, ki bo zajela maksimalno odstopanje v vrednostih. Druga komponenta bo tista, ki zajame maksimalno preostalo varianco in je hkrati pravokotna na prvo komponento. Podobno velja za ostale komponente, pri čemer velja omeniti, da so vse osnovne komponente med sabo pravokotne (neodvisne).

Čeprav lahko definiramo za m-dimenzionalne podatke m osnovnih komponent, se pogosto izkaže, da lahko večino informacij dobimo že z uporabo zelo majhnega števila osnovnih komponent. Kljub temu je pri transformaciji prostora z metodo *PCA* ob gradnji ansamblov klasifikatorjev potrebna previdnost. Osnovni cilj metode je ustvarjanje nove, manjše množice nekorelirajočih atributov, ki maksimizirajo ohranjanje variance podatkov. Pri tem se lahko pripeti, da izgubimo tiste komponente, katerih doprinos k variabilnosti podatkov je majhen, vendar so ključnega pomena za razločevanje med posameznimi vzorci (Fukunaga & Koontz, 1970). Zaradi tega se pri gradnji rotacijskega drevesa ohranijo vse komponente in se ne izvaja redukcija dimenzij.

Postopek iskanja osnovnih komponent je naslednji. Imejmo k m-dimenzionalnih primerov p_i in naj bo njihova srednja vrednost $\mu = 0$:

$$p_i = [a_1, a_2, \dots, a_m]^T \; ; \; i = 1 \dots k.$$

Če posamezne primere p_i združimo v matriko $X = [p_1, p_2, \dots, p_k]$, potem lahko za matriko X definiramo kovariančno matriko C kot $C = XX^T$. Za matriko C lahko izračunamo lastne vrednosti λ_i ter pripadajoče lastne vektorje v_i. Lastni vektor, ki pripada največji lastni vrednosti, predstavlja prvo osnovno komponento, lastni vektor, ki pripada drugi največji lastni vrednosti, predstavlja drugo osnovno komponento in tako naprej.

Predpostavka, ki smo jo uporabili pri računanju lastnih vektorjev, je, da je srednja vrednost primerov μ = 0. V primeru, ko to ne drži, definiramo

$$p_i' = p_i - \mu \; ; \; i = 1 \ldots k$$

in nato za spremenjene primere p_i' i izračunamo kovariančno matriko C in njene lastne vektorje.

Namen transformacije prostora z metodo *PCA* pri gradnji rotacijskega gozda ni zgolj morebitno izboljšanje diskriminatorske sposobnosti atributov, temveč povečanje raznolikosti posameznih osnovnih klasifikatorjev. Ukrep, ki dodatno pripomore k temu cilju, je tudi naključna izbira podmnožice vrednosti odločitvenega atributa, še preden se naključno izberejo vzorci iz osnovne učne množice. Šele nato se izvede transformacija s pomočjo metode *PCA*. Na ta način se dodatno zmanjša možnost pojavitve identičnih koeficientov v primerih, ko sta izbrani enaki podmnožici atributov.

4.1.2 Prilagoditev ansambla

Za razliko od naključnih gozdov lahko idejo rotacijskega gozda prenesemo tudi na druge tipe baznih metod, kar smo s pridom izkoristili. Za naš namen reševanja problema napovedovanja numeričnih vrednosti bi lahko uporabili kateregakoli izmed regresorjev, predstavljenih v 2. poglavju. Ker so študije pokazale, da so regresijska drevesa zaradi svoje občutljivosti na majhne spremembe v učni množici in posledično velike raznolikosti zgrajenih regresorjev zelo primerna za gradnjo ansamblov (Polikar, 2006), smo izbrali kar njih in vpeljali poimenovanje »rotacijski regresijski gozd«. Ustrezno alternativo bi lahko predstavljale tudi nevronske mreže, vendar smo zaradi prikazanih dobrih rezultatov na področju klasifikacije (Rodríguez, Kuncheva, & Alonso, 2006) ostali pri drevesih.

Na podlagi začetnih poskusov, ki so pri napovedovanju manjkajočih vrednosti pokazali večjo natančnost modelnih regresijskih dreves v primerjavi z regresijskimi drevesi, katerih listi vsebujejo konstantne vrednosti, smo se odločili za to varianto. Spreminjanje števila osnovnih regresijskih dreves je vplivalo na natančnost ansambla, vendar so višje vrednosti tega parametra (nad 10 dreves) privedle do zanemarljivega doprinosa k splošni natančnosti v primerjavi s povečano računsko zahtevnostjo. Poizkusili smo tudi s spreminjanjem drugih privzetih vrednosti parametrov rotacijskega gozda, vendar nismo dosegli izboljšanja rezultatov, zato smo obdržali začetne vrednosti. Nekoliko presenetljivo smo

dosegli višjo natančnost napovedi, če osnovna modelna drevesa po gradnji niso bila rezana. Predpostavljamo, da je to povzročilo večjo raznolikost osnovnih dreves.

gradnja rotacijskega regresijskega gozda:

VHOD: osnovna učna množica $\boldsymbol{X}$ z $\boldsymbol{N}$ vzorci, predstavljenimi z $\boldsymbol{n}$ atributi, množica vrednosti razrednega atributa $\boldsymbol{Y}$, število osnovnih regresijskih dreves $\boldsymbol{L}$, število podmnožic atributov $\boldsymbol{K}$

ZA $\boldsymbol{i}$ = 1 do $\boldsymbol{L}$:

- pripravi rotacijsko matriko $\boldsymbol{R_i^a}$:
  - razdeli množico vseh atributov $\boldsymbol{F}$ na $\boldsymbol{K}$ podmnožic $\boldsymbol{F_{i,j}}$ ($\boldsymbol{j}$ = 1 ... $\boldsymbol{K}$)
  - ZA $\boldsymbol{j}$ = 1 do $\boldsymbol{K}$:
    - naj bo $\boldsymbol{X_{i,j}}$ množica vzorcev iz $\boldsymbol{X}$ z atributi iz $\boldsymbol{F_{i,j}}$
    - iz $\boldsymbol{X_{i,j}}$ odstrani vse vzorce z vrednostmi razrednega atributa iz naključno izbrane podmnožice vseh vrednosti razrednega atributa
    - iz $\boldsymbol{X_{i,j}}$ z nadomeščanjem izberi 75% vzorcev in tako dobljeno množico označi z $\boldsymbol{X'_{i,j}}$
    - nad $\boldsymbol{X'_{i,j}}$ izvedi transformacijo z metodo *PCA* in pridobi koeficiente glavnih komponent v matriki $\boldsymbol{C_{i,j}}$
  - vstavi matrike $\boldsymbol{C_{i,j}}$ ($\boldsymbol{j}$ = 1 ... $\boldsymbol{K}$) v rotacijsko matriko $\boldsymbol{R_i}$, kot v formuli (4.1)
  - ustvari matriko $\boldsymbol{R_i^a}$ s preurejanjem stolpcev matrike $\boldsymbol{R_i}$, tako da ustrezajo zaporedju atributov v $\boldsymbol{F}$
- izgradi regresijsko drevo $\boldsymbol{D_i}$ na podlagi učne množice $\boldsymbol{XR_i^a}$

napoved vrednosti:

- za dani vzorec $\boldsymbol{x}$ naj bo $\boldsymbol{d_i(xR_i^a)}$ napovedana vrednosti, ki jo napove regresijsko drevo $\boldsymbol{D_i}$
- napovedana vrednost razrednega atributa vzorca $\boldsymbol{x}$ je aritmetična sredina napovedi vseh regresijskih dreves $\boldsymbol{D_i}$ ($\boldsymbol{i}$ = 1 ... $\boldsymbol{L}$):

$$\bar{r}(x) = \frac{1}{L}\sum_{i=1}^{L} d_i(xR_i^a)$$

Slika 4.1-1: Psevdokod algoritma gradnje rotacijskega regresijskega gozda

4.2 Stohastična metoda za izboljšanje ohranjanja variance

Vsaka napovedana numerična vrednost, ki jo na podlagi napovedi posameznih regresijskih dreves ustvari rotacijski regresijski gozd, je izračunana kot aritmetična sredina teh napovedi. Odvisno od posameznega primera se osnovna regresijska drevesa med seboj bolj ali manj strinjajo glede končne napovedi. Če izračunamo varianco osnovnih napovedi, lahko govorimo o ravni zaupanja q, ki je obratno sorazmerna izračunani varianci:

$$q = \frac{1}{\frac{1}{n}\sum_{k=1}^{n}(r^{(k)} - \bar{r})^2}, \tag{4.3}$$

kjer je $\bar{r}$ končna napovedana vrednost in $r^{(k)}$vrednost, ki jo napove k-to modelno drevo (k = 1,...,n). Nižja je raven zaupanja, večja je varianca napovedanih vrednosti, oz. posamezna drevesa so bolj nesoglasna.

Stohastična metoda za ohranjanje variance, ki smo jo razvili, temelji na ideji, da lahko izboljšamo ohranjanje variance s prilagajanjem končnih napovedi manjkajočih vrednosti, tako da ustrezne spremembe le-teh povzročijo spremembo skupne variance vseh vrednosti izbranega atributa, oz. njeno približevanje izhodiščni varianci pred nadomeščanjem manjkajočih vrednosti. Pri tem se večje spremembe izvajajo na tistih napovedih, pri katerih je raven zaupanja q manjša.

Razvili smo dve različici metode za ohranjanje variance, ki se razlikujeta v obsegu, v katerem spreminjata prvotne napovedane vrednosti. Neagresivna varianta metode opravi le manjše korekcije napovedanih vrednosti, medtem ko si agresivna različica dovoli večje spremembe. Obe varianti sta podrobneje opisani v nadaljevanju.

4.2.1 Neagresivna varianta

Osnovni namen neagresivne različice metode za ohranjanje variance je ovrednotiti, v kolikšni meri se lahko izboljša ohranjanje variance ob majhnih spremembah prvotnih napovedanih vrednosti in posledično tudi majhni spremembi splošne natančnosti rotacijskega regresijskega gozda.

S tem razlogom smo omejili modifikacijo vsake posamezne napovedi na ozek interval med aritmetičnim povprečjem vrednosti, ki jo za to manjkajočo vrednost napovejo osnovna regresijska

drevesa in obteženim povprečjem teh vrednosti $\bar{o}$, ki se po svoji velikosti nahaja med geometričnim in harmoničnim povprečjem:

$$\bar{o} = \sum_{k=1}^{n} \frac{\left(\frac{r^{(k)}}{\sum_{i=1}^{n} 1/d(r^{(i)})} \right)}{d(r^{(k)})}, \tag{4.4}$$

kjer je

$$d\left(r^{(k)}\right) = \sum_{i=1}^{n} \left| r^{(i)} - r^{(k)} \right|.$$

Na ta način smo omilil vpliv močno odstopajočih napovedi (osamelcev). Metoda za vsako napovedano vrednost izračuna njeno odstopanje od povprečne vrednosti atributa in če to odstopanje znaša več kot prednastavljeni prag, spremeni vrednost, tako da ji prišteje ali odšteje utež, ki je po velikosti sorazmerna varianci napovedane vrednosti. Celoten postopek se vrti v zanki, dokler se splošna varianca vseh vrednosti atributa razlikuje od začetne variance vrednosti tega atributa pred nadomeščanjem manjkajočih vrednosti za več kot prednastavljen prag, oziroma dokler nove napovedane vrednosti limitirajo k obteženim povprečjem $\bar{o}$ (Slika 4.2-1).

4.2.2 Agresivna varianta

Druga, agresivnejša varianta metode modificira napovedane vrednosti znotraj intervala med obema skrajnostma standardnega odklona vseh napovedi posameznih regresijskih dreves. Od neagresivne različice se torej razlikuje samo v obsegu, v katerem spreminja začetne napovedane vrednosti. Pri agresivni varianti ni potrebe po računanju obteženega povprečja. Doseže lahko precej bolj občutno ohranjanje variance v primerjavi z neagresivno različico, vendar je tudi vpliv na natančnost celotnega ansambla bolj konkreten. Medtem ko lahko pričakujemo, da so razlike v natančnosti med rotacijskim regresijskim gozdom brez dodatne metode za ohranjanje variance in ansamblom z neagresivno različico metode zanemarljive, pri agresivni varianti temu ni tako in se natančnost lahko občutno izboljša ali poslabša.

Neagresivna metoda za ohranjanje variance:

VHOD: osnovna učna množica $\boldsymbol{X}$ z $\boldsymbol{N}$ vzorci, množica $\boldsymbol{Y}$ z $\boldsymbol{M}$ vzorci, pri katerih je vrednosti razrednega atributa napovedal rotacijski regresijski gozd ($\boldsymbol{M} \leq \boldsymbol{N}$), prag $\boldsymbol{p}$

ZA vsak vzorec $\boldsymbol{j}$ iz $\boldsymbol{Y}$ ($\boldsymbol{j}$ = 1 do $\boldsymbol{M}$):

- izračunaj obteženo povprečje napovedi $\overline{\boldsymbol{o}}_{\boldsymbol{j}}$ po formuli (4.4)
- izračunaj varianco $\boldsymbol{v}_{\boldsymbol{j}}$ napovedi posameznih regresijskih dreves za razredni atribut vzorca

IZRAČUNAJ originalno varianco $\boldsymbol{v}_{\boldsymbol{o}}$ vrednosti razrednega atributa vseh vzorcev iz $\boldsymbol{X}$

IZRAČUNAJ varianco po nadomeščanju $\boldsymbol{v}_{\boldsymbol{n}}$ vrednosti razrednega atributa vseh vzorcev iz $\boldsymbol{X}$ in $\boldsymbol{Y}$

IZRAČUNAJ povprečno vrednost $\overline{\boldsymbol{x}}$ razrednega atributa vseh vzorcev iz $\boldsymbol{X}$

DOKLER je (ABS($\boldsymbol{v}_{\boldsymbol{o}} - \boldsymbol{v}_{\boldsymbol{n}}$) > $\boldsymbol{p}$ IN se ABS($\boldsymbol{v}_{\boldsymbol{o}} - \boldsymbol{v}_{\boldsymbol{n}}$) zmanjšuje) PONAVLJAJ:

- utež $\boldsymbol{u}$ = (ABS($\boldsymbol{v}_{\boldsymbol{o}} - \boldsymbol{v}_{\boldsymbol{n}}$))$^{1/2}$
- ZA vsako napovedano vrednost $\boldsymbol{x}_{\boldsymbol{j}}$ razrednega atributa vzorca $\boldsymbol{j}$ iz $\boldsymbol{Y}$ ($\boldsymbol{j}$ = 1 do $\boldsymbol{M}$):
  - ČE je ($\boldsymbol{v}_{\boldsymbol{n}} \leq \boldsymbol{v}_{\boldsymbol{o}}$ IN $\boldsymbol{x}_{\boldsymbol{j}} > \overline{\boldsymbol{x}}$) ALI ($\boldsymbol{v}_{\boldsymbol{n}} > \boldsymbol{v}_{\boldsymbol{o}}$ IN $\boldsymbol{x}_{\boldsymbol{j}} \leq \overline{\boldsymbol{x}}$):
    - $\boldsymbol{x}_{\boldsymbol{j}} = \boldsymbol{x}_{\boldsymbol{j}} + \boldsymbol{u} * \boldsymbol{v}_{\boldsymbol{j}}$
    - ČE je $\boldsymbol{x}_{\boldsymbol{j}} > \overline{\boldsymbol{o}}_{\boldsymbol{j}}$ POTEM $\boldsymbol{x}_{\boldsymbol{j}} = \overline{\boldsymbol{o}}_{\boldsymbol{j}}$
  - DRUGAČE ČE je ($\boldsymbol{v}_{\boldsymbol{n}} \leq \boldsymbol{v}_{\boldsymbol{o}}$ IN $\boldsymbol{x}_{\boldsymbol{j}} \leq \overline{\boldsymbol{x}}$) ALI ($\boldsymbol{v}_{\boldsymbol{n}} > \boldsymbol{v}_{\boldsymbol{o}}$ IN $\boldsymbol{x}_{\boldsymbol{j}} > \overline{\boldsymbol{x}}$):
    - $\boldsymbol{x}_{\boldsymbol{j}} = \boldsymbol{x}_{\boldsymbol{j}} - \boldsymbol{u} * \boldsymbol{v}_{\boldsymbol{j}}$
    - ČE je $\boldsymbol{x}_{\boldsymbol{j}} < \overline{\boldsymbol{o}}_{\boldsymbol{j}}$ POTEM $\boldsymbol{x}_{\boldsymbol{j}} = \overline{\boldsymbol{o}}_{\boldsymbol{j}}$
- Izračunaj novo varianco po nadomeščanju $\boldsymbol{v}_{\boldsymbol{n}}$ vrednosti razrednega atributa vseh vzorcev iz $\boldsymbol{X}$ in $\boldsymbol{Y}$

VRNI modificirano množico vzorcev $\boldsymbol{Y}$

Slika 4.2-1: Psevdokod neagresivne metode za ohranjanje variance

5 Implementacija in eksperimentalno okolje

5.1 Implementacija

Pri razvoju rotacijskega regresijskega gozda in celotnega eksperimentalnega okolja smo se opirali na obstoječe knjižnice odprtokodnega ogrodja za strojno učenje Weka, kar nam je prihranilo dodatno implementacijo že obstoječih algoritmov, ki smo jih uporabili pri naših eksperimentih.

Celoten projekt je bil implementiran v programskem jeziku Java, pri čemer smo za lažje delo uporabili odprtokodno razvojno okolje Eclipse, različica 3.5.1. Programska oprema, ki smo jo razvili, se sestoji iz 3 povezanih modulov:

- računski modul, znotraj katerega je implementiran naš ansambel,
- modul za pretvorbo različnih tekstovnih formatov za predstavitev podatkovnih množic v format ARFF,
- grafični uporabniški vmesnik (SWT).

Orodje tako uporabniku omogoča pripravo podatkovne množice skupaj z osnovno deskriptivno statistiko (število vzorcev in atributov, odstotek manjkajočih vrednosti, standardni odkloni, variance, itd.). Osnovno nalogo seveda predstavlja vstavljanje manjkajočih vrednosti, pri čemer je poleg našega ansambla na voljo še 9 drugih metod, katerih učinkovitost lahko primerjamo. Naš namen je bil v primerjavo vključiti raznovrstne metode, ki se danes najpogosteje uporabljajo za nadomeščanje manjkajočih vrednosti. Večina jih je že implementiranih v odprtokodni knjižnici ogrodja za strojno učenje Weka, dodali pa smo tudi metodo orodja za statistično analizo SPSS, pri čemer smo nadomeščanje manjkajočih vrednosti izvedli znotraj programskega paketa SPSS in v naše orodje uvozili izhodne datoteke (dopolnjene podatkovne množice). Metode, ki jih je možno vključiti v primerjavo, so:

- regresijsko drevo,
- modelno drevo,
- metoda najmanjših srednjih kvadratov,
- linearna regresija,
- metoda k-najbližjih sosedov,

- vstavljanje povprečne vrednosti,
- vstavljanje ničel,
- bagging,
- metoda večkratnega vstavljanja orodja SPSS,
- rotacijski regresijski gozd.

Metode so podrobneje predstavljene v poglavjih 2, 3 in 4. Pri nadomeščanju manjkajočih vrednosti s pomočjo rotacijskega regresijskega gozda je možno uporabiti tudi dodatni metodi za ohranjanje variance, opisani v prejšnjem poglavju.

Uporabnik lahko spreminja parametre posameznih metod, iz postopka izključi izbrane atribute podatkovne množice in pred nadomeščanjem izvede transformacijo numeričnih vrednosti (na voljo sta normalizacija in logaritemska pretvorba).

Orodje za vsako vključeno metodo ustvari novo izhodno podatkovno množico, ki vsebuje vstavljene nadomestne vrednosti. Učinkovito ovrednotenje posameznih metod je možno le, če so na voljo dejanske originalne vrednosti. Manjkajoče vrednosti se zato lahko naključno generirajo, pri čemer uporabnik določi njihov odstotek in izbere atribute, ki naj jih vsebujejo. Na podlagi primerjave nadomestnih vrednosti z originalnimi nam računski modul vrne rezultate v obliki naslednjih metrik:

- povprečna kvadratna napaka,
- koren povprečne kvadratne napake,
- povprečna absolutna napaka,
- relativna kvadratna napaka,
- relativna absolutna napaka,
- koeficient korelacije,
- varianca pred in po vstavljanju.

Posamezne metrike so opisane v 2. poglavju.

Za potrebe simulacije manjkajočih vrednosti tipa *MAR* in *NMAR* smo implementirali tudi metodo, ki ustvari podatkovno množico iz naključno generiranih numeričnih vrednosti, pri čemer lahko uporabnik določi dimenzije množice in parametre odvisnosti razrednega atributa od preostalih, neodvisnih atributov.

5.2 Eksperimentalno okolje

V tem razdelku so predstavljene podatkovne množice, ki smo jih izbrali za izvajanje naših meritev, ter protokoli izvajanje le-teh.

5.2.1 Podatkovne množice

Za učinkovito in celostno ovrednotenje napovedovanja manjkajočih numeričnih vrednosti s pomočjo rotacijskega regresijskega gozda in preostalih metod, ki smo jih vključili v primerjavo z našim ansamblom, smo morali izbrati zadostno število dovolj raznolikih podatkovnih množic s pretežno numeričnimi atributi. Na njih smo izvajali poizkuse, ki so obsegali glavnino našega dela, osredotočenega na manjkajoče vrednosti tipa *MCAR*, pri katerih smo si obetali najboljše rezultate. Mehanizma nastanka manjkajočih vrednosti *MAR* in *NMAR* smo simulirali z generiranjem umetne podatkovne množice in ustrezno selekcijo vrednosti, ki smo jih naknadno iz množice odstranili.

Javno dostopne množice

Obstoječe podatkovne množice smo izbrali predvsem iz javno dostopnega vira podatkovnih množic, namenjenih strojnemu učenju: UCI Machine Laerning Repository (Asuncion & Newman, 2007). Izjema sta le podatkovni množici »Bohen« in »Spellman«, ki vsebujeta izmerjene genske ekspresije ter smo jih v raziskavo vključili zaradi primerjave z rezultati obstoječe študije (Brock, Shaffer, Blakesley, Lotz, & Tseng, 2008). Poskušali smo izbrati čim bolj raznolike množice z različnimi stopnjami korelacije med posameznimi atributi in porazdelitvami njihovih vrednosti. Izbrali smo 15 podatkovnih množic, vendar smo množico »*Musk*« izključili iz selekcije, saj je le-ta zaradi svojih velikih dimenzij in velikega števila načrtovanih meritev predstavljala časovno pretrd oreh. Osnovne značilnosti izbranih podatkovnih množic so prikazane v spodnji tabeli.

Tabela 5.2-1: Osnovne lastnosti izbranih javno dostopnih podatkovnih množic

Ime množice	Število vseh atributov	Število numeričnih atributov	Število vzorcev	Odstotek manjkajočih vrednosti	Opomba
Bohen	16	16	16523	7,1 %	genske ekspresije
Concrete	9	9	1030	0 %	
Cpu	7	7	209	0 %	večina atributov vsebuje intervalne vrednosti
E.coli	9	7	336	0 %	
Japanese vowels	12	12	9961	0 %	združeni 2 podatkovni množici
Lrs	103	102	531	0 %	
Mammographic masses	6	6	961	2,8 %	intervalne in ordinalne vrednosti
Ozone	73	72	2536	8,3 %	
Page blocks	11	11	5473	0 %	1 atribut z ordinalnimi vrednostmi
Pima indians – diabetes	9	9	768	0 %	1 binarni nominalni atribut
Segmentation	20	19	210	0 %	
Spellman	82	77	6178	5,9 %	genske ekspresije
Wisconsin breast cancer	31	30	568	0 %	
Yeast	10	8	1484	0 %	

Javno dostopne množice so bile shranjene v različnih podatkovnih formatih. Pred izvajanjem meritev smo jih pretvorili v format ARFF, ki je bil predstavljen v 2. poglavju.

Umetno generirana podatkovna množica

Kadar poskušamo z metodami strojnega učenja napovedati numerično vrednost na podlagi znanja, pridobljenega skozi učni proces, domnevamo, da med napovedano vrednostjo in preostalimi (znanimi) vrednostmi obstaja določena relacija oziroma odvisnost. Mehanizma nastanka naključnih (*MAR*) in nenaključnih (*NMAR*) manjkajočih vrednosti sicer nista nujno pogojena z obstojem takšne odvisnosti, vendar se posledice razlik med mehanizmi nastanka manjkajočih vrednosti pokažejo v celoti šele takrat, ko so vrednosti razrednega atributa odvisne od vrednosti enega ali več drugih atributov pripadajočega vzorca v podatkovni množici. Zaradi tega se pri primerjavi metod za nadomeščanje manjkajočih vrednosti, nastalih po mehanizmih *MAR* in *NMAR*, nismo zanašali na podatkovne množice, na katerih smo izvajali poskuse z manjkajočimi vrednostmi tipa *MCAR*. Da smo namreč lahko zagotovili zahtevano odvisnost med atributi, smo ustvarili lastno podatkovno množico s tremi neodvisnimi atributi in odvisnim, razrednim atributom. Vrednosti razrednega atributa te množice smo izračunali na podlagi poljubno generiranih vrednosti preostalih 3 atributov vsakega posameznega vzorca, in sicer kot rezultat multivariatnega polinoma tretje stopnje z naključno določenimi koeficienti:

$$w = a_3x^3 + a_2y^2 + a_1z + a_0 + e, \qquad (5.1)$$

kjer so x, y in z vrednosti neodvisnih atributov in w vrednost razrednega atributa posameznega vzorca podatkovne množice. Koeficient e je variabilen in je bil za vsaki vzorec množice naključno izbran iz intervala [-0,01, 0,01], s čimer smo vpeljali napako in zmanjšali »nerealno« odvisnost razrednega atributa. Posamezne vrednosti neodvisnih atributov kakor tudi preostali koeficienti so bili naključno izbrani iz intervala [-0,5, 0,5]. Na ta način smo ustvarili 1000 različnih vzorcev.

5.2.2 Protokol izvajanja meritev

Zaradi že opisanega vpliva različnih mehanizmov nastanka manjkajočih vrednosti (*MCAR*, *MAR* in *NMAR*) na kasnejšo sposobnost napovedovanja le-teh, smo posebej obravnavali vsako izmed treh možnih situacij.

Manjkajoče vrednosti, katerih mehanizem nastanka je popolnoma naključen (*Missing completely at random*), smo simulirali z naključnim vstavljanjem v 14 različnih podatkovnih množic, ki smo jih (z izjemo dveh podatkovnih baz z genskimi ekspresijami) pridobili iz spletne shrambe *UCI Machine Learning Repository*. Za preostala mehanizma nastanka manjkajočih vrednosti (*MAR* in *NMAR*) smo uporabili zgoraj opisano lastno podatkovno množico, pri kateri smo lahko zagotovili potrebno odvisnost nastanka manjkajočih vrednosti od vrednosti atributov množice.

Osnovni cilj naših preverjanj je bil ugotavljanje natančnosti naše metode za nadomeščanje manjkajočih vrednosti, ki smo jo primerjali z 9 drugimi metodami. Za oceno natančnosti posameznih metod smo uporabili metriko »koren povprečne kvadratne napake« (*RMSE*, glej formulo 2.3), ki smo jo za potrebe eksperimenta priredili za preverjanje natančnosti na vseh atributih podatkovne množice. To skupno oceno natančnosti za celotno podatkovno množico smo izračunali po naslednji formuli:

$$TRMSE = \sqrt{\frac{\sum_{i=0}^{k} SE_i}{N}}, \tag{5.2}$$

kjer je N število vseh manjkajočih vrednosti v podatkovni množici in

$$SE_i = (p_1^{(i)} - a_1^{(i)})^2 + \cdots + (p_{n^{(i)}}^{(i)} - a_{n^{(i)}}^{(i)})^2 \tag{5.3}$$

ter je $n^{(i)}$ število manjkajočih vrednosti atributa i, $p_{k^{(i)}}^{(i)}$ in $a_{k^{(i)}}^{(i)}$ pa sta napovedana k-ta manjkajoča vrednost atributa i in njena dejanska vrednost. Razširitev metrike napake na vse atribute podatkovne množice ima za posledico občutljivost na razliko v rangu velikosti vrednosti posameznih atributov. Zaradi tega smo pred izvajanjem naših poskusov normalizirali vrednosti vseh atributov naših podatkovnih množic, torej smo vse vrednosti posameznega numeričnega atributa proporcionalno preslikali v vrednosti iz intervala [0, 1].

Poleg merjenja natančnosti smo za vsako metodo ocenjevali tudi sposobnost ohranjanja variance vrednosti posameznih atributov po vstavljanju manjkajočih vrednosti. Pomen ohranjanja variance je bil opisan v 3. in 4. poglavju, kjer smo tudi predstavili dve različici stohastične metode za izboljšanje ohranjanja variance ob uporabi z rotacijskim regresijskim gozdom. Varianco vrednosti posameznega atributa smo izračunali po formuli za varianco vrednosti končne množice X z n numeričnimi elementi:

$$Var(X) = \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n}. \tag{5.4}$$

6 Rezultati

V tem poglavju predstavljamo rezultate empiričnih preverjanj, na podlagi katerih smo se odločili, ali lahko sprejmemo ali zavrnemo vsako izmed posameznih hipotez, ki smo jih postavili v uvodnem poglavju.

6.1 Učinkovitost metod pri mehanizmu MCAR

Mehanizem nastanka povsem naključnih manjkajočih vrednosti smo simulirali z vstavljanjem manjkajočih vrednosti v 14 različnih podatkovnih množic, iz katerih smo predhodno odstranili nepopolne vzorce. Eksperiment smo ponovili za 7 različnih stopenj manjkajočih vrednosti, tako da smo iz vsake množice naključno odstranili 1%, 5%, 10%, 15%, 20%, 25% in 50% vseh vrednosti (Slika 6.1-1). Vstavljene manjkajoče vrednosti so bile enakomerno porazdeljene po vseh atributih posamezne množice. Ker nismo mogli uporabiti prečnega preverjanja, smo za vsako stopnjo manjkajočih vrednosti izvedli 5 poskusov, tako do smo ustvarili po 5 različnih množic z enako stopnjo manjkajočih vrednosti. Skupaj smo tako opravili 490 poskusov (5 ponovitev za vsako izmed 7 stopenj manjkajočih vrednosti na 14 podatkovnih množicah).

Pri vsakem poskusu smo preverjali natančnost 10 metod za napovedovanje manjkajočih vrednosti, pri čemer smo naš rotacijski regresijski gozd uporabili tudi v kombinaciji z dvema različicama metode za izboljšanje ohranjanja variance, ki smo jo razvili. Tako je skupno število vseh metod, ki smo jih primerjali, zraslo na 12. Poleg treh različic našega ansambla smo uporabili navadno regresijsko drevo, modelno regresijsko drevo, linearno regresijo, metodo najmanjših povprečnih kvadratov, metodo k-najbližjih sosedov, nadomeščanje z vrednostjo 0, nadomeščanje s povprečno vrednostjo, metodo *bagging* in metodo večkratnega vstavljanja, ki jo ponuja orodje za statistično analizo SPSS. Vsakič smo izračunali skupno napako uporabljene metode kot aritmetično povprečno vrednost napak *TRMSE*, ki jih je metoda ustvarila pri nadomeščanju manjkajočih vrednosti. Rezultati so prikazani v spodnjih tabelah in grafih.

VHOD: odstotek manjkajočih vrednosti ***o***, podatkovna množica ***M***

PREŠTEJ število vzorcev ***n*** in število vseh atributov ***m*** v podatkovni množici ***M***

IZRAČUNAJ število vseh manjkajočih vrednosti ***mv***:

mv = ***n*** * ***m*** * ***o*** / 100

PONAVLJAJ:

PONAVLJAJ:

naključno IZBERI atribut ***a*** iz podatkovne množice ***M***

DOKLER ni ***a.število_manjkajočih_vrednosti*** < 1,25 * ***n*** *** o*** / 100

PONAVLJAJ:

naključno IZBERI vzorec ***v*** iz podatkovne množice ***M***

DOKLER je ***v.a*** manjkajoča vrednost

IZBRIŠI ***v.a***

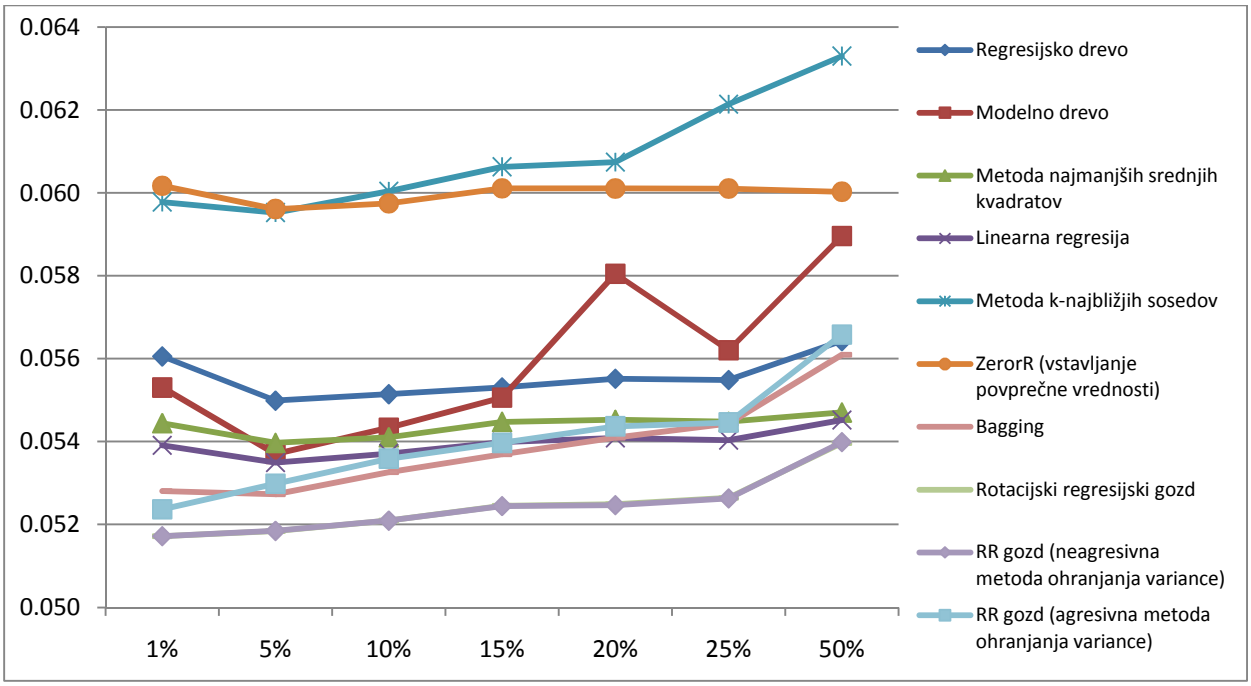
DOKLER število vstavljenih manjkajočih vrednosti ni enako ***mv***

Slika 6.1-1: Psevdokod algoritma za vstavljanje manjkajočih vrednosti

Tabela 6.1-1: Povprečne napake po metriki TRMSE na podatkovni množici *Bohen*

Metoda*	Bohen (1%)	Bohen (5%)	Bohen (10%)	Bohen (15%)	Bohen (20%)	Bohen (25%)	Bohen (50%)
Reg. drevo	5,61E-2 ± 1,30E-3	5,50E-2 ± 1,26E-3	5,51E-2 ± 4,74E-4	5,53E-2 ± 4,35E-4	5,55E-2 ± 3,47E-4	5,55E-2 ± 1,27E-4	5,64E-2 ± 3,68E-4
Modelno drevo	5,53E-2 ± 1,31E-3	5,37E-2 ± 1,80E-3	5,43E-2 ± 1,24E-3	5,51E-2 ± 9,48E-4	5,81E-2 ± 4,28E-3	5,62E-2 ± 2,01E-3	5,90E-2 ± 3,41E-3
MNSK	5,44E-2 ± 9,47E-4	5,40E-2 ± 1,21E-3	5,41E-2 ± 3,26E-4	5,45E-2 ± 4,79E-4	5,45E-2 ± 4,35E-4	5,45E-2 ± 3,89E-4	5,47E-2 ± 2,62E-4
Lin. Reg.	5,39E-2 ± 9,86E-4	5,35E-2 ± 1,11E-3	5,37E-2 ± 2,95E-4	5,40E-2 ± 5,16E-4	5,41E-2 ± 3,62E-4	5,40E-2 ± 3,81E-4	5,45E-2 ± 2,55E-4
K-NN	5,98E-2 ± 1,23E-3	5,95E-2 ± 9,13E-4	6,00E-2 ± 7,59E-4	6,06E-2 ± 6,69E-4	6,07E-2 ± 6,74E-4	6,21E-2 ± 7,76E-4	6,33E-2 ± 1,55E-3
ZeroR	6,02E-2 ± 1,25E-3	5,96E-2 ± 1,24E-3	5,97E-2 ± 4,08E-4	6,01E-2 ± 4,69E-4	6,01E-2 ± 3,65E-4	6,01E-2 ± 3,25E-4	6,00E-2 ± 1,53E-4
Vstavi 0	5,33E-1 ± 3,64E-3	5,31E-1 ± 6,53E-4	5,32E-1 ± 6,04E-4	5,34E-1 ± 4,44E-3	5,32E-1 ± 6,70E-4	5,32E-1 ± 6,72E-4	5,31E-1 ± 3,04E-4
Bagging	5,28E-2 ± 8,28E-4	5,27E-2 ± 1,22E-3	5,33E-2 ± 2,73E-4	5,37E-2 ± 4,68E-4	5,41E-2 ± 3,86E-4	5,44E-2 ± 3,24E-4	5,61E-2 ± 1,79E-4
Rot. reg. gozd	5,17E-2 ± 1,09E-3	5,18E-2 ± 1,05E-3	5,21E-2 ± 3,63E-4	5,25E-2 ± 8,19E-4	5,25E-2 ± 4,61E-4	5,26E-2 ± 1,52E-4	5,40E-2 ± 3,56E-4
RRG var. 1	5,17E-2 ± 1,09E-3	5,18E-2 ± 1,05E-3	5,21E-2 ± 3,83E-4	5,24E-2 ± 8,49E-4	5,25E-2 ± 4,84E-4	5,26E-2 ± 1,43E-4	5,40E-2 ± 3,76E-4
RRG var. 2	5,24E-2 ± 9,49E-4	5,30E-2 ± 1,03E-3	5,36E-2 ± 2,51E-4	5,40E-2 ± 5,14E-4	5,44E-2 ± 2,24E-4	5,45E-2 ± 3,30E-4	5,66E-2 ± 3,58E-4
SPSS	7,63E-2 ± 4,36E-4	7,67E-2 ± 2,89E-4	7,62E-2 ± 1,01E-3	7,71E-2 ± 1,62E-4	7,73E-2 ± 4,36E-5	7,81E-2 ± 2,21E-4	8,08E-2 ± 1,38E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

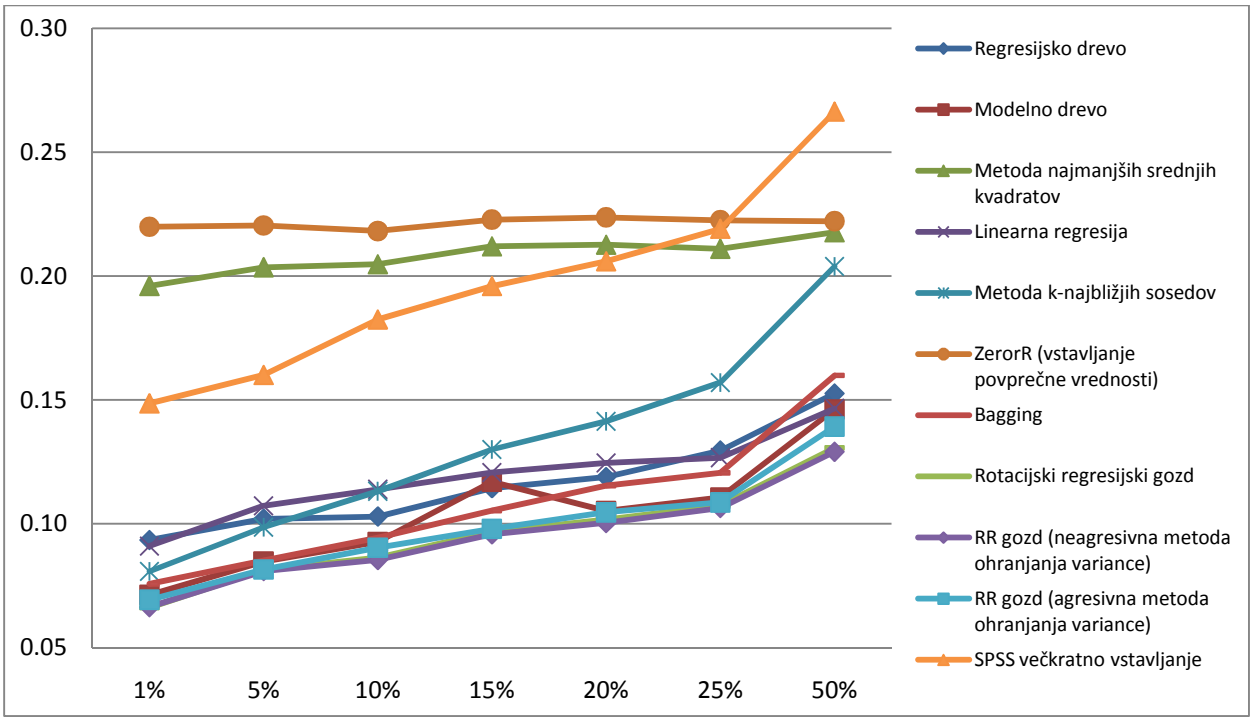


Graf 6.1-1: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Bohen* (metoda orodja za statistično analizo SPSS in metoda vstavljanja ničel zaradi preglednosti grafa nista vključeni, saj njune povprečne napake močno odstopajo)

Tabela 6.1-2: Povprečne napake po metriki TRMSE na podatkovni množici *Concrete*

Metoda*	Concrete (1%)	Concrete (5%)	Concrete (10%)	Concrete (15%)	Concrete (20%)	Concrete (25%)	Concrete (50%)
Reg. drevo	9,34E-2 ± 1,56E-2	1,02E-1 ± 1,30E-2	1,03E-1 ± 1,33E-2	1,14E-1 ± 7,46E-3	1,19E-1 ± 1,32E-2	1,30E-1 ± 7,26E-3	1,53E-1 ± 2,51E-3
Modelno drevo	7,13E-2 ± 5,68E-3	8,48E-2 ± 5,35E-3	9,27E-2 ± 1,24E-2	1,17E-1 ± 3,63E-2	1,05E-1 ± 8,98E-3	1,11E-1 ± 3,95E-3	1,46E-1 ± 8,43E-3
MNSK	1,96E-1 ± 1,65E-2	2,04E-1 ± 1,18E-2	2,05E-1 ± 1,08E-2	2,12E-1 ± 5,49E-3	2,13E-1 ± 4,55E-3	2,11E-1 ± 4,92E-3	2,18E-1 ± 2,56E-3
Lin. Reg.	9,09E-2 ± 4,04E-3	1,07E-1 ± 5,68E-3	1,14E-1 ± 7,59E-3	1,21E-1 ± 1,89E-3	1,25E-1 ± 2,31E-3	1,27E-1 ± 1,98E-3	1,47E-1 ± 3,76E-3
K-NN	8,08E-2 ± 9,66E-3	9,86E-2 ± 3,38E-3	1,13E-1 ± 5,71E-3	1,30E-1 ± 4,13E-3	1,41E-1 ± 6,38E-3	1,57E-1 ± 3,95E-3	2,04E-1 ± 7,02E-3
ZeroR	2,20E-1 ± 1,21E-2	2,20E-1 ± 1,04E-2	2,18E-1 ± 8,58E-3	2,23E-1 ± 3,18E-3	2,24E-1 ± 3,25E-3	2,23E-1 ± 6,98E-4	2,22E-1 ± 1,12E-3
Vstavi 0	4,21E-1 ± 2,72E-2	4,24E-1 ± 1,29E-2	4,23E-1 ± 8,68E-3	4,27E-1 ± 5,09E-3	4,28E-1 ± 7,16E-3	4,27E-1 ± 2,99E-3	4,27E-1 ± 1,71E-3
Bagging	7,57E-2 ± 9,39E-3	8,52E-2 ± 8,70E-3	9,43E-2 ± 8,40E-3	1,05E-1 ± 2,00E-3	1,15E-1 ± 4,44E-3	1,21E-1 ± 2,09E-3	1,60E-1 ± 2,82E-3
Rot. reg. gozd	6,61E-2 ± 3,85E-3	8,14E-2 ± 5,55E-3	8,63E-2 ± 1,22E-2	9,68E-2 ± 6,36E-3	1,02E-1 ± 9,58E-3	1,07E-1 ± 5,34E-3	1,31E-1 ± 5,42E-3
RRG var. 2	6,64E-2 ± 3,99E-3	8,09E-2 ± 5,59E-3	8,54E-2 ± 1,28E-2	9,58E-2 ± 6,95E-3	1,00E-1 ± 1,03E-2	1,06E-1 ± 5,86E-3	1,29E-1 ± 5,62E-3
RRG var. 2	6,93E-2 ± 4,04E-3	8,16E-2 ± 2,66E-3	9,03E-2 ± 6,89E-3	9,79E-2 ± 3,43E-3	1,05E-1 ± 4,86E-3	1,09E-1 ± 3,59E-3	1,39E-1 ± 4,54E-3
SPSS	1,49E-1 ± 1,84E-3	1,60E-1 ± 2,52E-3	1,82E-1 ± 6,64E-3	1,96E-1 ± 1,13E-2	2,06E-1 ± 5,22E-3	2,19E-1 ± 2,72E-3	2,66E-1 ± 2,08E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

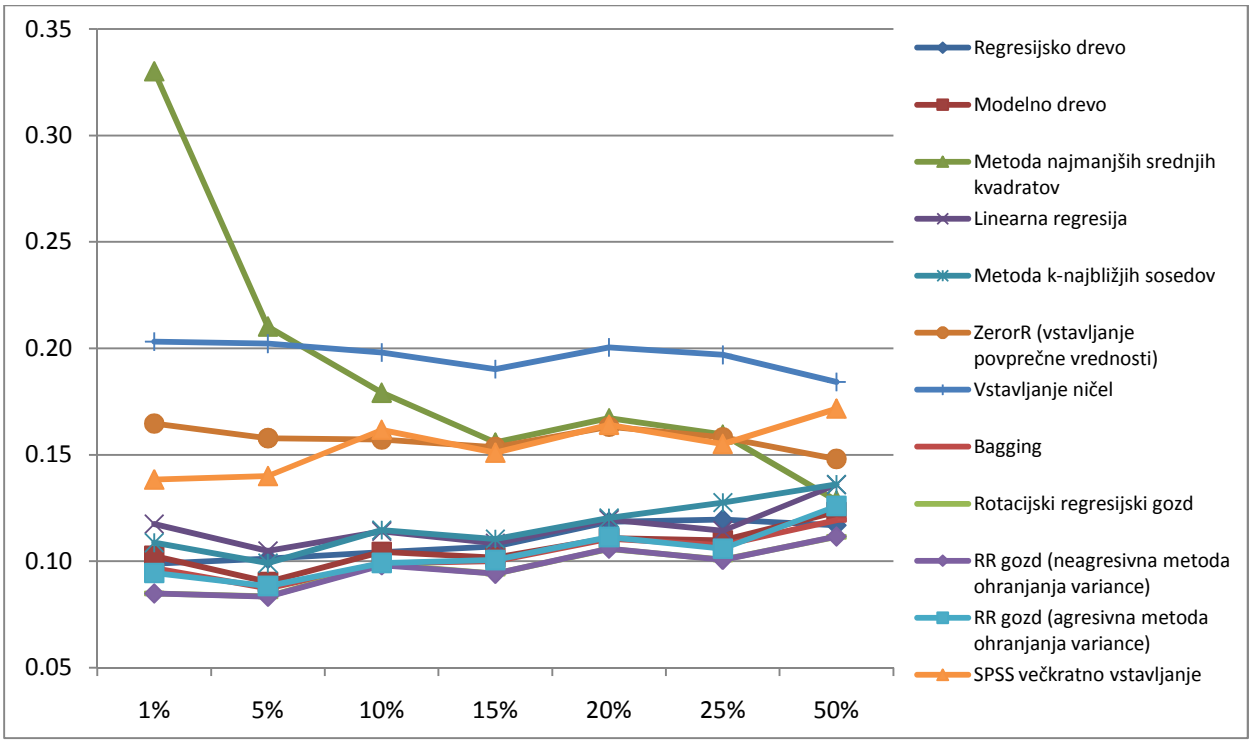


Graf 6.1-2: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Concrete* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-3: Povprečne napake po metriki TRMSE na podatkovni množici *Cpu*

Metoda*	Cpu (1%)	Cpu (5%)	Cpu (10%)	Cpu (15%)	Cpu (20%)	Cpu (25%)	Cpu (50%)
Reg. drevo	9,88E-2 ± 4,32E-2	1,01E-1 ± 2,53E-2	1,04E-1 ± 1,12E-2	1,07E-1 ± 1,69E-2	1,18E-1 ± 6,81E-3	1,20E-1 ± 1,78E-2	1,17E-1 ± 1,77E-3
Modelno drevo	1,02E-1 ± 4,42E-2	9,03E-2 ± 2,18E-2	1,04E-1 ± 2,24E-2	1,02E-1 ± 1,68E-2	1,11E-1 ± 1,22E-2	1,10E-1 ± 1,39E-2	1,23E-1 ± 6,67E-3
MNSK	3,30E-1 ± 4,00E-1	2,10E-1 ± 7,91E-2	1,79E-1 ± 5,29E-2	1,56E-1 ± 2,87E-2	1,67E-1 ± 4,80E-2	1,60E-1 ± 2,32E-2	1,28E-1 ± 1,15E-2
Lin. Reg.	1,17E-1 ± 5,05E-2	1,05E-1 ± 2,02E-2	1,14E-1 ± 1,62E-2	1,08E-1 ± 9,21E-3	1,20E-1 ± 1,23E-2	1,14E-1 ± 9,07E-3	1,36E-1 ± 1,36E-2
K-NN	1,09E-1 ± 5,41E-2	9,91E-2 ± 1,82E-2	1,15E-1 ± 1,81E-2	1,10E-1 ± 9,11E-3	1,20E-1 ± 1,45E-2	1,28E-1 ± 1,41E-2	1,36E-1 ± 1,12E-2
ZeroR	1,65E-1 ± 5,57E-2	1,58E-1 ± 1,73E-2	1,57E-1 ± 1,97E-2	1,54E-1 ± 1,09E-2	1,63E-1 ± 1,71E-2	1,58E-1 ± 8,70E-3	1,48E-1 ± 9,36E-3
Vstavi 0	2,03E-1 ± 7,61E-2	2,02E-1 ± 2,39E-2	1,98E-1 ± 2,24E-2	1,90E-1 ± 1,44E-2	2,00E-1 ± 2,04E-2	1,97E-1 ± 9,97E-3	1,84E-1 ± 1,05E-2
Bagging	9,67E-2 ± 5,06E-2	8,73E-2 ± 2,17E-2	9,90E-2 ± 1,21E-2	1,00E-1 ± 7,64E-3	1,11E-1 ± 1,29E-2	1,08E-1 ± 1,12E-2	1,19E-1 ± 1,02E-2
Rot. reg. gozd	8,49E-2 ± 4,17E-2	8,35E-2 ± 2,04E-2	9,85E-2 ± 2,08E-2	9,42E-2 ± 9,00E-3	1,06E-1 ± 1,19E-2	1,01E-1 ± 7,29E-3	1,11E-1 ± 5,86E-3
RRG var. 2	8,48E-2 ± 4,19E-2	8,33E-2 ± 2,07E-2	9,81E-2 ± 2,08E-2	9,41E-2 ± 8,95E-3	1,06E-1 ± 1,18E-2	1,01E-1 ± 7,41E-3	1,12E-1 ± 6,00E-3
RRG var. 3	9,44E-2 ± 4,73E-2	8,84E-2 ± 2,01E-2	9,92E-2 ± 1,75E-2	1,01E-1 ± 8,94E-3	1,11E-1 ± 1,37E-2	1,06E-1 ± 3,93E-3	1,26E-1 ± 9,47E-3
SPSS	1,38E-1 ± 5,06E-2	1,40E-1 ± 8,94E-3	1,62E-1 ± 3,84E-3	1,51E-1 ± 1,03E-2	1,64E-1 ± 9,91E-3	1,55E-1 ± 1,55E-2	1,72E-1 ± 1,53E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

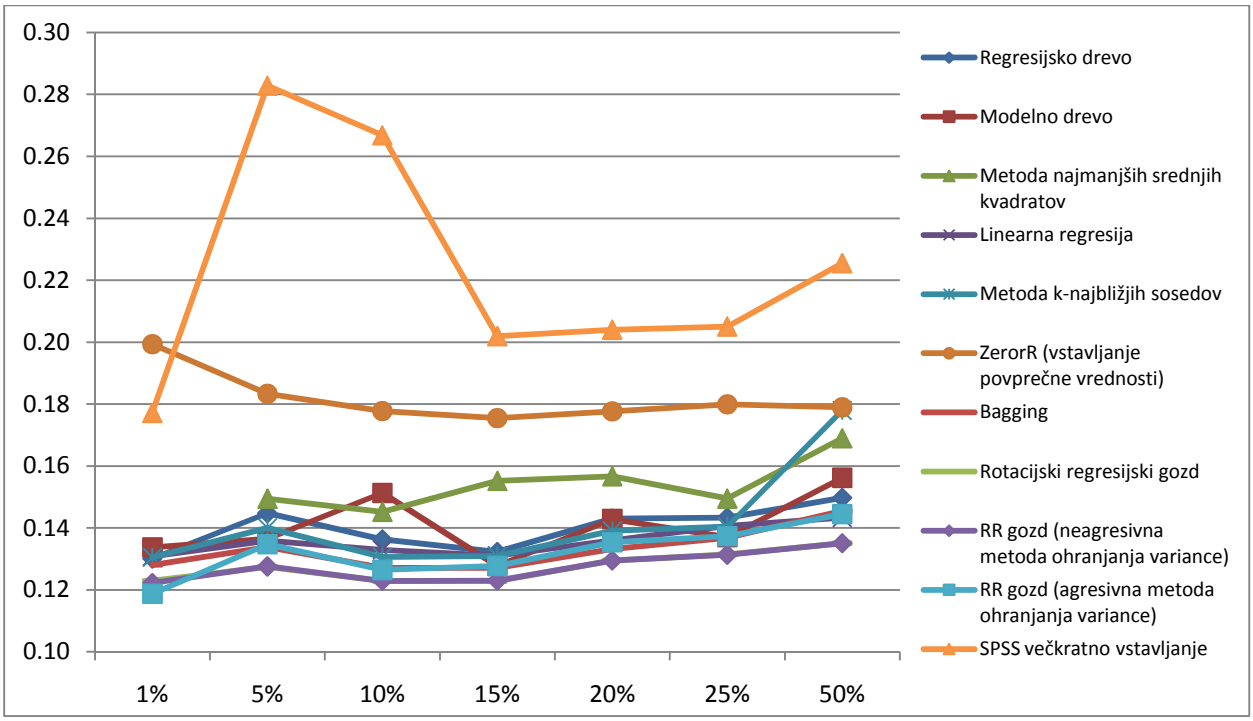


Graf 6.1-3: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Cpu*

Tabela 6.1-4: Povprečne napake po metriki TRMSE na podatkovni množici *E. coli*

Metoda*	E. coli (1%)	E. coli (5%)	E. coli (10%)	E. coli (15%)	E. coli (20%)	E. coli (25%)	E. coli (50%)
Reg. drevo	1,30E-1 ± 5,47E-2	1,45E-1 ± 2,56E-2	1,36E-1 ± 1,50E-2	1,32E-1 ± 8,30E-3	1,43E-1 ± 1,10E-2	1,43E-1 ± 1,99E-3	1,50E-1 ± 7,15E-3
Modelno drevo	1,34E-1 ± 4,76E-2	1,37E-1 ± 2,55E-2	1,51E-1 ± 3,14E-2	1,28E-1 ± 6,53E-3	1,43E-1 ± 2,36E-2	1,37E-1 ± 5,68E-3	1,56E-1 ± 3,98E-2
MNSK	1,47E-1 ± 5,47E-2	1,49E-1 ± 1,66E-2	1,45E-1 ± 1,74E-2	1,55E-1 ± 3,15E-2	1,57E-1 ± 2,22E-2	1,50E-1 ± 5,79E-3	1,69E-1 ± 3,04E-2
Lin. Reg.	1,31E-1 ± 5,04E-2	1,36E-1 ± 2,29E-2	1,33E-1 ± 1,54E-2	1,31E-1 ± 5,20E-3	1,36E-1 ± 1,24E-2	1,41E-1 ± 3,84E-3	1,43E-1 ± 7,88E-3
K-NN	1,31E-1 ± 5,40E-2	1,40E-1 ± 2,60E-2	1,31E-1 ± 1,73E-2	1,31E-1 ± 7,22E-3	1,39E-1 ± 8,41E-3	1,40E-1 ± 3,19E-3	1,78E-1 ± 1,76E-2
ZeroR	1,99E-1 ± 2,30E-2	1,83E-1 ± 1,91E-2	1,78E-1 ± 1,09E-2	1,75E-1 ± 9,67E-3	1,78E-1 ± 1,04E-2	1,80E-1 ± 2,88E-3	1,79E-1 ± 6,04E-3
Vstavi 0	4,89E-1 ± 5,09E-2	4,56E-1 ± 2,09E-2	4,70E-1 ± 2,59E-2	4,56E-1 ± 1,44E-2	4,65E-1 ± 5,72E-3	4,62E-1 ± 7,54E-3	4,66E-1 ± 6,76E-3
Bagging	1,28E-1 ± 4,70E-2	1,34E-1 ± 2,31E-2	1,27E-1 ± 1,47E-2	1,27E-1 ± 5,97E-3	1,33E-1 ± 1,28E-2	1,37E-1 ± 4,21E-3	1,45E-1 ± 7,49E-3
Rot. reg. gozd	1,23E-1 ± 4,85E-2	1,27E-1 ± 2,46E-2	1,23E-1 ± 1,62E-2	1,23E-1 ± 5,89E-3	1,30E-1 ± 1,30E-2	1,31E-1 ± 3,64E-3	1,35E-1 ± 7,09E-3
RRG var. 2	1,22E-1 ± 4,84E-2	1,28E-1 ± 2,48E-2	1,23E-1 ± 1,62E-2	1,23E-1 ± 5,70E-3	1,30E-1 ± 1,31E-2	1,31E-1 ± 3,79E-3	1,35E-1 ± 7,14E-3
RRG var. 2	1,19E-1 ± 4,64E-2	1,35E-1 ± 2,01E-2	1,26E-1 ± 1,41E-2	1,28E-1 ± 5,76E-3	1,35E-1 ± 1,35E-2	1,37E-1 ± 4,71E-3	1,44E-1 ± 6,88E-3
SPSS	1,77E-1 ± 6,75E-2	2,83E-1 ± 1,05E-1	2,67E-1 ± 9,21E-2	2,02E-1 ± 5,66E-3	2,04E-1 ± 5,84E-4	2,05E-1 ± 4,12E-4	2,25E-1 ± 8,03E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

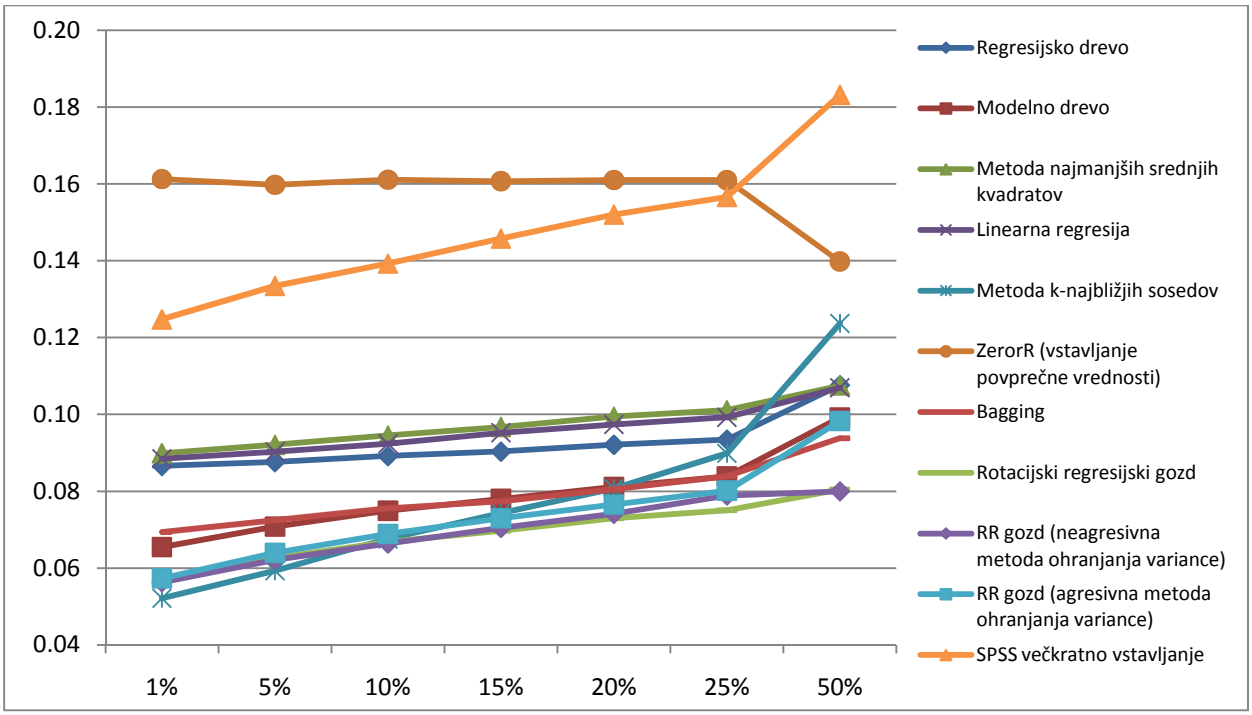


Graf 6.1-4: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *E. coli* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-5: Povprečne napake po metriki TRMSE na podatkovni množici *Japanese vowels*

Metoda*	J. vowels (1%)	J. vowels (5%)	J. vowels (10%)	J. vowels (15%)	J. vowels (20%)	J. vowels (25%)	J. vowels (50%)
Reg. drevo	8,66E-2 ± 3,30E-3	8,76E-2 ± 7,90E-4	8,92E-2 ± 4,51E-4	9,03E-2 ± 1,21E-3	9,21E-2 ± 1,73E-3	9,34E-2 ± 1,91E-3	1,08E-1 ± 2,42E-3
Modelno drevo	6,55E-2 ± 9,79E-4	7,08E-2 ± 7,30E-4	7,49E-2 ± 8,05E-4	7,80E-2 ± 1,14E-3	8,11E-2 ± 1,37E-3	8,39E-2 ± 1,26E-3	9,92E-2 ± 3,11E-3
MNSK	8,99E-2 ± 2,56E-3	9,21E-2 ± 4,57E-4	9,45E-2 ± 6,72E-4	9,67E-2 ± 4,29E-4	9,94E-2 ± 7,12E-4	1,01E-1 ± 4,27E-4	1,07E-1 ± 7,58E-3
Lin. Reg.	8,85E-2 ± 2,20E-3	9,02E-2 ± 4,85E-4	9,24E-2 ± 5,86E-4	9,52E-2 ± 4,20E-4	9,74E-2 ± 3,47E-4	9,93E-2 ± 3,97E-4	1,07E-1 ± 4,93E-3
K-NN	5,21E-2 ± 1,12E-3	5,93E-2 ± 5,01E-4	6,76E-2 ± 7,34E-4	7,43E-2 ± 8,56E-4	8,07E-2 ± 2,84E-4	8,99E-2 ± 8,15E-4	1,24E-1 ± 3,04E-2
ZeroR	1,61E-1 ± 1,53E-3	1,60E-1 ± 2,03E-3	1,61E-1 ± 7,22E-4	1,61E-1 ± 5,63E-4	1,61E-1 ± 1,03E-3	1,61E-1 ± 7,90E-4	1,40E-1 ± 4,72E-2
Vstavi 0	5,23E-1 ± 5,32E-3	5,23E-1 ± 1,15E-3	5,23E-1 ± 9,37E-4	5,22E-1 ± 6,56E-4	5,23E-1 ± 2,53E-4	5,23E-1 ± 9,48E-4	4,29E-1 ± 2,11E-1
Bagging	6,94E-2 ± 2,06E-3	7,25E-2 ± 4,21E-4	7,56E-2 ± 1,47E-4	7,74E-2 ± 3,64E-3	8,06E-2 ± 3,82E-3	8,39E-2 ± 3,54E-3	9,38E-2 ± 2,72E-2
Rot. reg. gozd	5,64E-2 ± 1,95E-3	6,27E-2 ± 3,05E-4	6,68E-2 ± 4,37E-4	6,97E-2 ± 1,98E-3	7,29E-2 ± 2,34E-3	7,50E-2 ± 3,13E-3	8,05E-2 ± 2,13E-2
RRG var. 2	5,63E-2 ± 1,93E-3	6,22E-2 ± 3,63E-4	6,64E-2 ± 4,72E-4	7,05E-2 ± 3,83E-3	7,41E-2 ± 4,54E-3	7,89E-2 ± 1,15E-2	8,00E-2 ± 2,22E-2
RRG var. 2	5,73E-2 ± 1,64E-3	6,40E-2 ± 1,01E-3	6,89E-2 ± 9,39E-4	7,30E-2 ± 8,14E-4	7,65E-2 ± 4,35E-4	8,02E-2 ± 5,92E-4	9,83E-2 ± 4,01E-4
SPSS	1,25E-1 ± 9,03E-5	1,33E-1 ± 1,33E-3	1,39E-1 ± 4,62E-3	1,46E-1 ± 9,11E-4	1,52E-1 ± 1,35E-3	1,57E-1 ± 1,66E-4	1,83E-1 ± 5,81E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

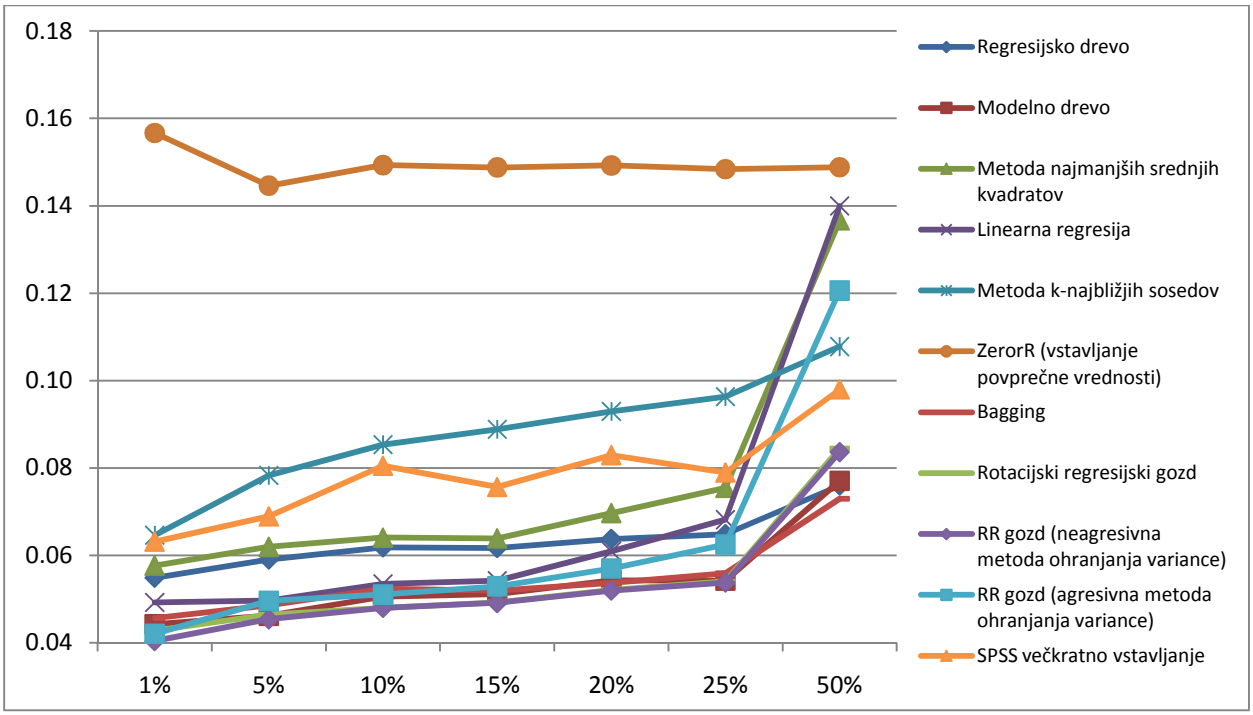


Graf 6.1-5: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Japanese vowels* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-6: Povprečne napake po metriki TRMSE na podatkovni množici *LRS*

Metoda*	Lrs (1%)	Lrs (5%)	Lrs (10%)	Lrs (15%)	Lrs (20%)	Lrs (25%)	Lrs (50%)
Reg. drevo	5,49E-2 ± 7,22E-3	5,91E-2 ± 3,60E-3	6,18E-2 ± 2,94E-3	6,17E-2 ± 3,49E-3	6,38E-2 ± 1,80E-3	6,49E-2 ± 1,83E-3	7,60E-2 ± 1,42E-3
Modelno drevo	4,43E-2 ± 7,10E-3	4,62E-2 ± 3,16E-3	5,06E-2 ± 4,25E-3	5,11E-2 ± 2,16E-3	5,43E-2 ± 1,90E-3	5,42E-2 ± 1,63E-3	7,71E-2 ± 1,59E-3
MNSK	5,77E-2 ± 9,45E-3	6,20E-2 ± 5,88E-3	6,41E-2 ± 2,78E-3	6,39E-2 ± 3,17E-3	6,97E-2 ± 2,07E-3	7,54E-2 ± 8,31E-4	1,37E-1 ± 2,94E-3
Lin. Reg.	4,92E-2 ± 1,08E-2	4,97E-2 ± 2,89E-3	5,35E-2 ± 3,87E-3	5,42E-2 ± 2,76E-3	6,10E-2 ± 2,08E-3	6,82E-2 ± 2,02E-3	1,40E-1 ± 2,81E-3
K-NN	6,46E-2 ± 4,22E-3	7,83E-2 ± 4,27E-3	8,53E-2 ± 1,89E-3	8,89E-2 ± 3,76E-3	9,30E-2 ± 3,04E-3	9,63E-2 ± 4,88E-3	1,08E-1 ± 5,38E-3
ZeroR	1,57E-1 ± 8,42E-3	1,45E-1 ± 2,70E-3	1,49E-1 ± 2,14E-3	1,49E-1 ± 2,36E-3	1,49E-1 ± 2,01E-3	1,48E-1 ± 9,30E-4	1,49E-1 ± 7,82E-4
Vstavi 0	4,61E-1 ± 1,83E-2	4,59E-1 ± 5,31E-3	4,59E-1 ± 4,40E-3	4,61E-1 ± 2,74E-3	4,60E-1 ± 1,27E-3	4,60E-1 ± 3,04E-3	4,60E-1 ± 4,10E-4
Bagging	4,56E-2 ± 4,24E-3	4,86E-2 ± 2,57E-3	5,24E-2 ± 3,37E-3	5,20E-2 ± 2,38E-3	5,37E-2 ± 1,50E-3	5,59E-2 ± 1,27E-3	7,30E-2 ± 1,99E-3
Rot. reg. gozd	4,28E-2 ± 7,39E-3	4,65E-2 ± 5,02E-3	4,80E-2 ± 4,05E-3	4,92E-2 ± 1,57E-3	5,21E-2 ± 1,75E-3	5,40E-2 ± 1,13E-3	8,45E-2 ± 2,10E-3
RRG var. 2	4,04E-2 ± 6,74E-3	4,54E-2 ± 2,93E-3	4,80E-2 ± 4,16E-3	4,92E-2 ± 1,54E-3	5,20E-2 ± 1,80E-3	5,38E-2 ± 1,17E-3	8,36E-2 ± 2,21E-3
RRG var. 2	4,21E-2 ± 5,69E-3	4,96E-2 ± 2,75E-3	5,10E-2 ± 4,71E-3	5,29E-2 ± 3,00E-3	5,70E-2 ± 7,43E-4	6,25E-2 ± 1,51E-3	1,21E-1 ± 9,63E-4
SPSS	6,32E-2 ± 5,58E-3	6,89E-2 ± 6,72E-3	8,05E-2 ± 4,06E-3	7,56E-2 ± 1,33E-3	8,29E-2 ± 5,69E-3	7,89E-2 ± 3,20E-4	9,80E-2 ± 8,41E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

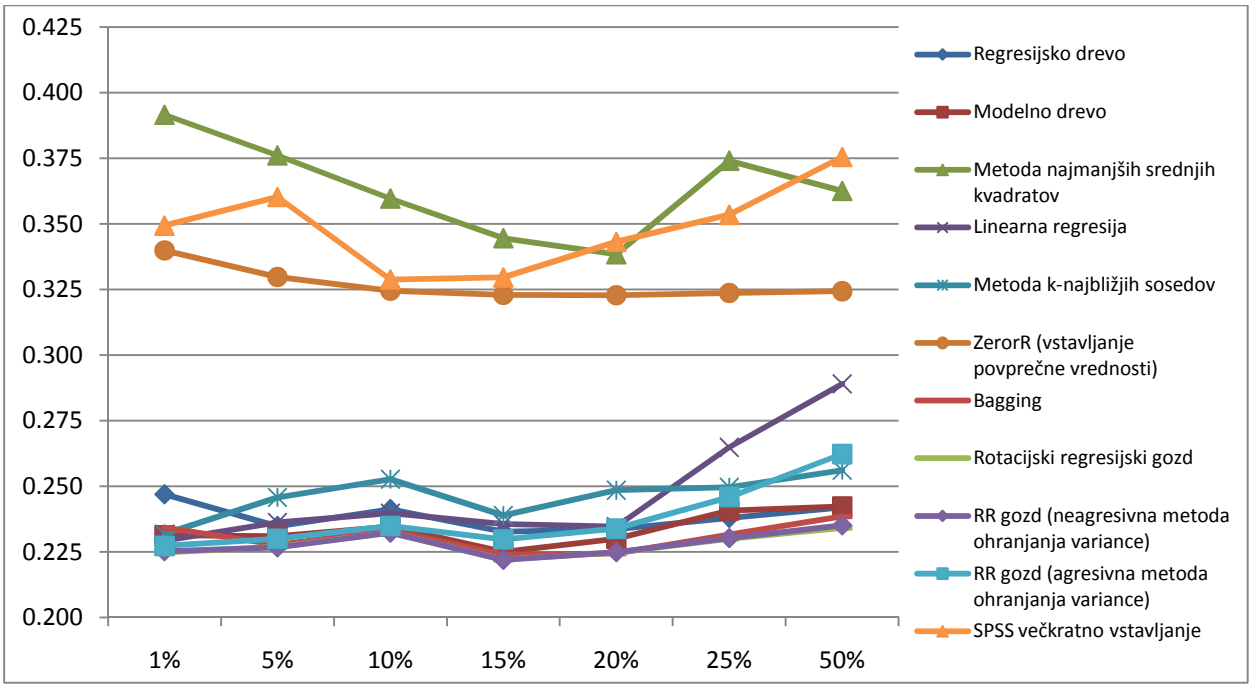


Graf 6.1-6: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *LRS* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-7: Povprečne napake po metriki TRMSE na podatkovni množici *Mammographic masses*

Metoda*	Mam. m. (1%)	Mam. m. (5%)	Mam. m. (10%)	Mam. m. (15%)	Mam. m. (20%)	Mam. m. (25%)	Mam. m. (50%)
Reg. drevo	2,47E-1 ± 2,78E-2	2,35E-1 ± 1,28E-2	2,41E-1 ± 1,18E-2	2,33E-1 ± 1,00E-2	2,34E-1 ± 9,85E-3	2,38E-1 ± 4,44E-3	2,42E-1 ± 3,67E-3
Modelno drevo	2,31E-1 ± 2,85E-2	2,31E-1 ± 1,79E-2	2,35E-1 ± 6,64E-3	2,25E-1 ± 6,64E-3	2,30E-1 ± 1,35E-2	2,41E-1 ± 1,38E-2	2,42E-1 ± 3,20E-3
MNSK	3,92E-1 ± 3,81E-2	3,76E-1 ± 4,62E-2	3,60E-1 ± 2,53E-2	3,45E-1 ± 2,13E-2	3,38E-1 ± 1,56E-2	3,74E-1 ± 4,05E-2	3,63E-1 ± 2,84E-2
Lin. Reg.	2,29E-1 ± 2,27E-2	2,36E-1 ± 1,83E-2	2,40E-1 ± 7,40E-3	2,36E-1 ± 1,21E-2	2,35E-1 ± 1,11E-2	2,65E-1 ± 1,46E-2	2,89E-1 ± 6,27E-2
K-NN	2,32E-1 ± 2,49E-2	2,46E-1 ± 2,00E-2	2,53E-1 ± 9,16E-3	2,39E-1 ± 1,07E-2	2,49E-1 ± 1,00E-2	2,50E-1 ± 9,42E-3	2,56E-1 ± 3,79E-3
ZeroR	3,40E-1 ± 1,14E-2	3,30E-1 ± 1,95E-2	3,25E-1 ± 6,67E-3	3,23E-1 ± 6,30E-3	3,23E-1 ± 4,29E-3	3,24E-1 ± 4,47E-3	3,24E-1 ± 2,90E-3
Vstavi 0	5,46E-1 ± 3,41E-2	5,92E-1 ± 2,17E-2	5,85E-1 ± 1,42E-2	5,72E-1 ± 8,64E-3	5,83E-1 ± 1,06E-2	5,87E-1 ± 8,33E-3	5,80E-1 ± 2,84E-3
Bagging	2,34E-1 ± 2,08E-2	2,28E-1 ± 1,18E-2	2,33E-1 ± 8,20E-3	2,24E-1 ± 6,42E-3	2,24E-1 ± 7,40E-3	2,32E-1 ± 2,67E-3	2,39E-1 ± 3,15E-3
Rot. reg. gozd	2,25E-1 ± 2,71E-2	2,27E-1 ± 1,51E-2	2,32E-1 ± 6,90E-3	2,22E-1 ± 5,78E-3	2,25E-1 ± 6,85E-3	2,30E-1 ± 2,53E-3	2,34E-1 ± 3,41E-3
RRG var. 2	2,25E-1 ± 2,73E-2	2,27E-1 ± 1,58E-2	2,32E-1 ± 6,79E-3	2,22E-1 ± 5,50E-3	2,25E-1 ± 6,94E-3	2,30E-1 ± 2,74E-3	2,35E-1 ± 2,66E-3
RRG var. 2	2,27E-1 ± 3,04E-2	2,30E-1 ± 1,98E-2	2,35E-1 ± 4,67E-3	2,30E-1 ± 1,50E-2	2,34E-1 ± 1,08E-2	2,46E-1 ± 5,10E-3	2,62E-1 ± 1,56E-2
SPSS	3,49E-1 ± 2,93E-2	3,60E-1 ± 1,35E-2	3,29E-1 ± 8,35E-3	3,30E-1 ± 6,11E-3	3,43E-1 ± 8,34E-3	3,54E-1 ± 9,47E-3	3,76E-1 ± 1,95E-2

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

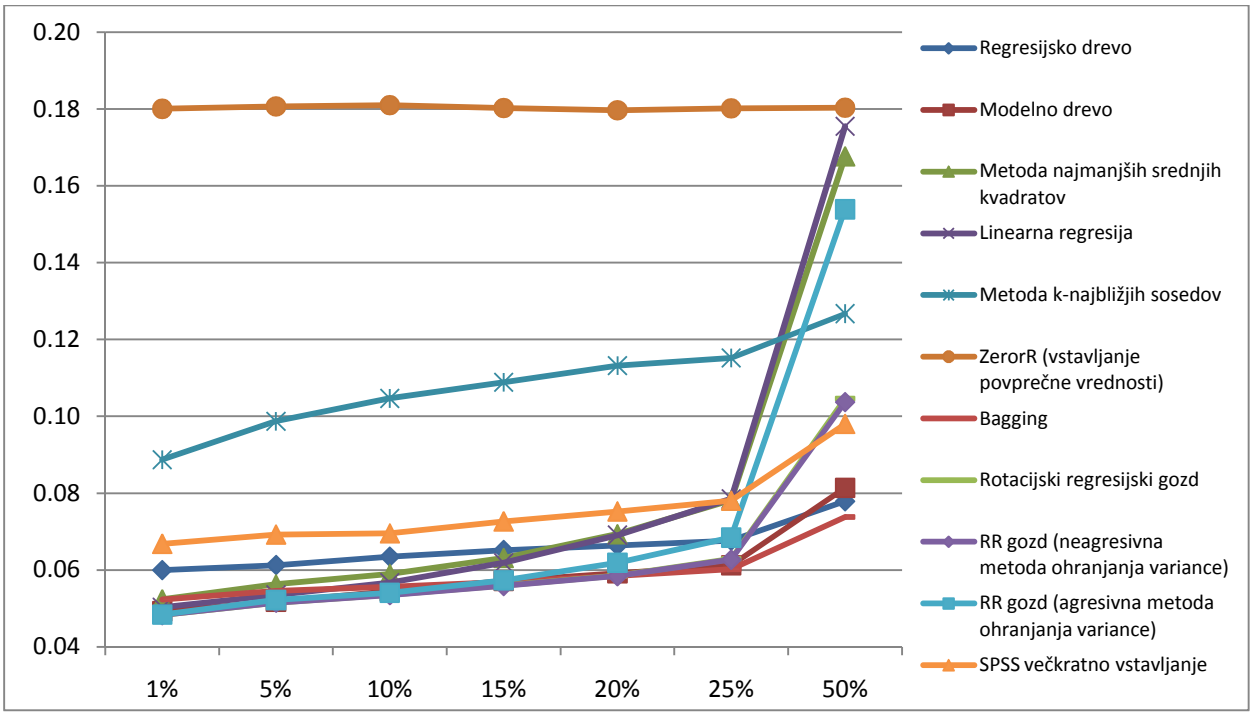


Graf 6.1-7: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Mammographic masses* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-8: Povprečne napake po metriki TRMSE na podatkovni množici *Ozone*

Metoda*	Ozone (1%)	Ozone (5%)	Ozone (10%)	Ozone (15%)	Ozone (20%)	Ozone (25%)	Ozone (50%)
Reg. drevo	5,99E-2 ± 2,14E-3	6,13E-2 ± 1,59E-3	6,35E-2 ± 1,14E-3	6,51E-2 ± 6,05E-4	6,64E-2 ± 7,73E-4	6,76E-2 ± 4,30E-4	7,79E-2 ± 1,29E-3
Modelno drevo	4,92E-2 ± 3,09E-3	5,17E-2 ± 1,38E-3	5,45E-2 ± 1,80E-3	5,71E-2 ± 8,42E-4	5,92E-2 ± 9,51E-4	6,12E-2 ± 1,07E-3	8,14E-2 ± 7,72E-3
MNSK	5,24E-2 ± 4,20E-3	5,63E-2 ± 2,30E-3	5,90E-2 ± 3,45E-4	6,32E-2 ± 5,48E-4	6,93E-2 ± 3,03E-4	7,82E-2 ± 1,36E-3	1,68E-1 ± 1,67E-3
Lin. Reg.	5,03E-2 ± 3,21E-3	5,35E-2 ± 1,22E-3	5,67E-2 ± 2,89E-4	6,18E-2 ± 3,76E-4	6,91E-2 ± 1,53E-4	7,86E-2 ± 1,17E-3	1,75E-1 ± 1,75E-3
K-NN	8,87E-2 ± 4,42E-3	9,87E-2 ± 9,31E-4	1,05E-1 ± 1,88E-3	1,09E-1 ± 2,61E-3	1,13E-1 ± 1,92E-3	1,15E-1 ± 3,74E-3	1,27E-1 ± 1,81E-3
ZeroR	1,80E-1 ± 2,02E-3	1,81E-1 ± 8,23E-4	1,81E-1 ± 1,18E-3	1,80E-1 ± 8,30E-4	1,80E-1 ± 4,63E-4	1,80E-1 ± 5,75E-4	1,80E-1 ± 3,41E-4
Vstavi 0	5,37E-1 ± 5,51E-3	5,32E-1 ± 1,61E-3	5,34E-1 ± 1,58E-3	5,34E-1 ± 8,55E-4	5,33E-1 ± 6,27E-4	5,33E-1 ± 1,20E-3	5,34E-1 ± 4,78E-4
Bagging	5,23E-2 ± 2,77E-3	5,45E-2 ± 1,55E-3	5,56E-2 ± 5,15E-4	5,69E-2 ± 6,44E-4	5,85E-2 ± 6,78E-4	6,02E-2 ± 5,87E-4	7,38E-2 ± 4,14E-4
Rot. reg. gozd	4,83E-2 ± 3,14E-3	5,15E-2 ± 1,31E-3	5,35E-2 ± 8,34E-4	5,59E-2 ± 7,22E-4	5,85E-2 ± 6,09E-4	6,29E-2 ± 2,31E-3	1,05E-1 ± 2,62E-2
RRG var. 2	4,83E-2 ± 3,12E-3	5,15E-2 ± 1,33E-3	5,35E-2 ± 8,21E-4	5,59E-2 ± 7,56E-4	5,84E-2 ± 6,53E-4	6,27E-2 ± 2,32E-3	1,04E-1 ± 2,64E-2
RRG var. 2	4,84E-2 ± 2,89E-3	5,22E-2 ± 1,33E-3	5,40E-2 ± 7,16E-4	5,73E-2 ± 3,11E-4	6,19E-2 ± 7,92E-4	6,84E-2 ± 1,07E-3	1,54E-1 ± 1,69E-3
SPSS	6,68E-2 ± 7,00E-4	6,92E-2 ± 2,24E-3	6,95E-2 ± 5,27E-4	7,26E-2 ± 2,88E-4	7,52E-2 ± 1,29E-3	7,80E-2 ± 1,51E-4	9,80E-2 ± 6,35E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

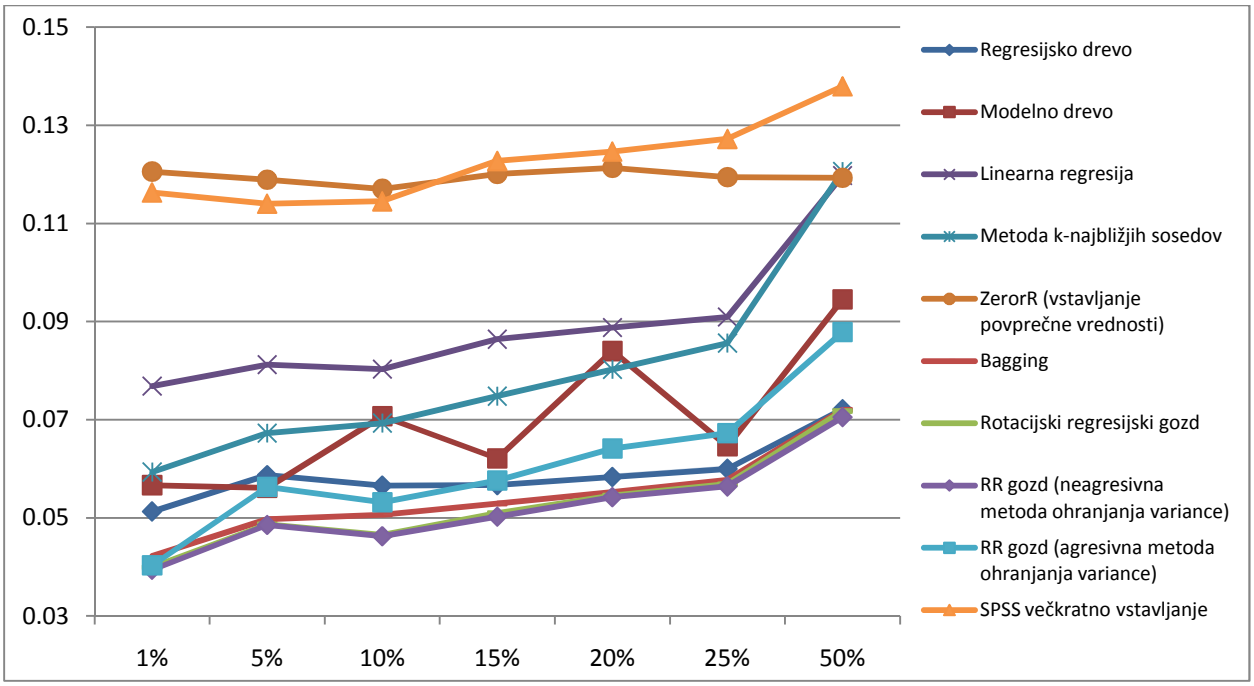


Graf 6.1-8: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Ozone* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-9: Povprečne napake po metriki TRMSE na podatkovni množici *Page blocks*

Metoda*	P. blocks (1%)	P. blocks (5%)	P. blocks (10%)	P. blocks (15%)	P. blocks (20%)	P. blocks (25%)	P. blocks (50%)
Reg. drevo	5,13E-2 ± 6,90E-3	5,87E-2 ± 5,44E-3	5,66E-2 ± 4,52E-3	5,67E-2 ± 1,15E-3	5,83E-2 ± 1,39E-3	6,00E-2 ± 2,62E-3	7,21E-2 ± 2,37E-3
Modelno drevo	5,66E-2 ± 2,01E-2	5,61E-2 ± 1,37E-2	7,07E-2 ± 3,71E-2	6,21E-2 ± 7,69E-3	8,40E-2 ± 6,05E-2	6,46E-2 ± 1,21E-2	9,45E-2 ± 8,23E-3
MNSK	3,99E-1 ± 3,39E-1	2,91E-1 ± 8,87E-2	3,80E-1 ± 1,83E-1	2,46E-1 ± 9,79E-2	3,15E-1 ± 1,79E-1	1,52E-1 ± 2,30E-2	2,17E-1 ± 1,51E-1
Lin. Reg.	7,68E-2 ± 1,14E-2	8,12E-2 ± 6,82E-3	8,03E-2 ± 1,33E-3	8,64E-2 ± 5,35E-3	8,88E-2 ± 6,66E-3	9,09E-2 ± 7,98E-3	1,20E-1 ± 7,82E-3
K-NN	5,93E-2 ± 1,02E-2	6,72E-2 ± 3,67E-3	6,93E-2 ± 2,62E-3	7,48E-2 ± 2,15E-3	8,02E-2 ± 3,06E-3	8,56E-2 ± 1,02E-3	1,21E-1 ± 5,93E-3
ZeroR	1,21E-1 ± 1,55E-2	1,19E-1 ± 6,90E-3	1,17E-1 ± 2,70E-3	1,20E-1 ± 5,74E-4	1,21E-1 ± 2,18E-3	1,19E-1 ± 1,39E-3	1,19E-1 ± 1,04E-3
Vstavi 0	2,93E-1 ± 1,35E-2	2,90E-1 ± 8,08E-3	2,82E-1 ± 1,73E-3	2,87E-1 ± 1,17E-3	2,85E-1 ± 2,09E-3	2,86E-1 ± 4,37E-3	2,85E-1 ± 1,42E-3
Bagging	4,22E-2 ± 9,15E-3	4,97E-2 ± 4,34E-3	5,06E-2 ± 2,59E-3	5,29E-2 ± 1,55E-3	5,53E-2 ± 8,47E-4	5,78E-2 ± 2,23E-3	7,20E-2 ± 1,13E-3
Rot. reg. gozd	4,00E-2 ± 8,55E-3	4,87E-2 ± 5,30E-3	4,65E-2 ± 2,81E-3	5,09E-2 ± 1,39E-3	5,46E-2 ± 9,57E-4	5,69E-2 ± 3,19E-3	7,18E-2 ± 1,55E-3
RRG var. 2	3,94E-2 ± 8,68E-3	4,85E-2 ± 5,31E-3	4,62E-2 ± 2,73E-3	5,02E-2 ± 1,02E-3	5,42E-2 ± 8,89E-4	5,64E-2 ± 3,33E-3	7,05E-2 ± 1,85E-3
RRG var. 2	4,03E-2 ± 1,03E-2	5,63E-2 ± 7,34E-3	5,32E-2 ± 3,35E-3	5,76E-2 ± 1,82E-3	6,41E-2 ± 2,99E-3	6,72E-2 ± 3,01E-3	8,79E-2 ± 5,59E-3
SPSS	1,16E-1 ± 7,28E-3	1,14E-1 ± 6,36E-3	1,15E-1 ± 1,18E-3	1,23E-1 ± 3,68E-3	1,25E-1 ± 4,79E-3	1,27E-1 ± 4,97E-4	1,38E-1 ± 6,24E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

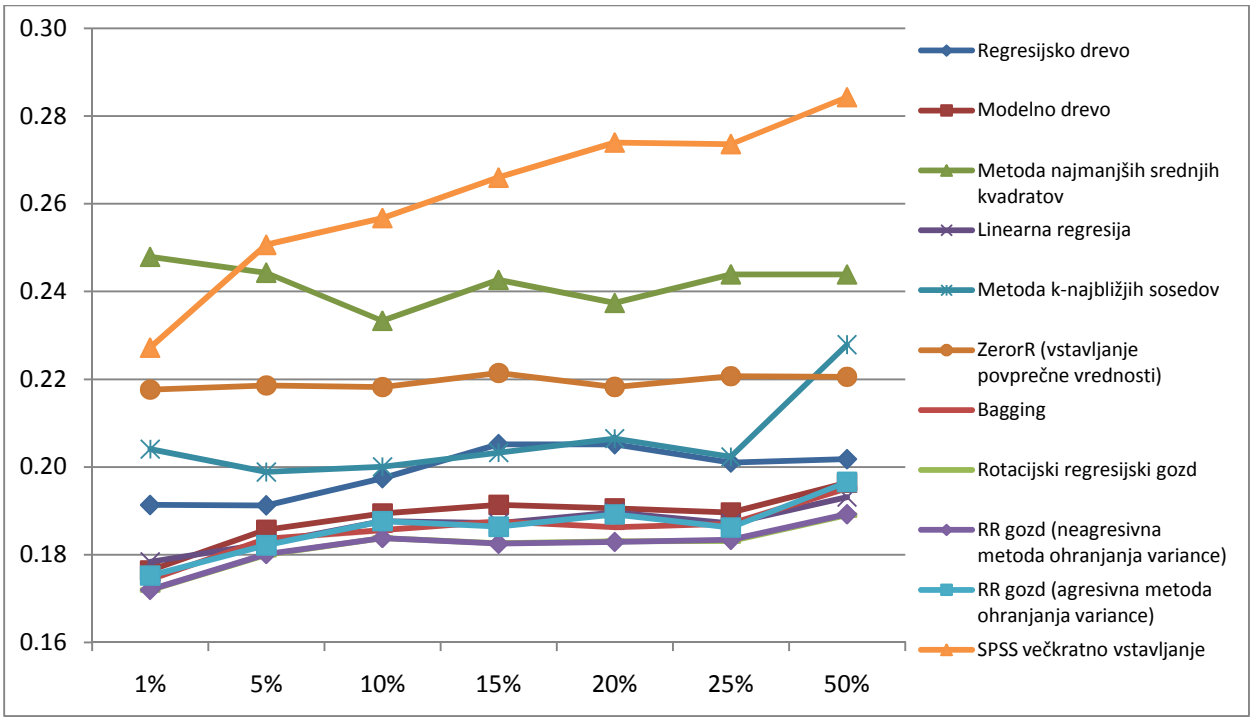


Graf 6.1-9: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Page blocks* (metoda najmanjših srednjih kvadratov in metoda vstavljanja ničel zaradi preglednosti grafa nista vključeni, saj njune povprečne napake močno odstopajo)

Tabela 6.1-10: Povprečne napake po metriki TRMSE na podatkovni množici *Pima Indians (diabetes)*

Metoda*	Pima Ind. (1%)	Pima Ind. (5%)	Pima Ind. (10%)	Pima Ind. (15%)	Pima Ind. (20%)	Pima Ind. (25%)	Pima Ind. (50%)
Reg. drevo	1,91E-1 ± 4,89E-2	1,91E-1 ± 1,32E-2	1,97E-1 ± 1,09E-2	2,05E-1 ± 1,47E-2	2,05E-1 ± 9,13E-3	2,01E-1 ± 9,16E-3	2,02E-1 ± 4,22E-3
Modelno drevo	1,76E-1 ± 3,46E-2	1,86E-1 ± 1,04E-2	1,89E-1 ± 1,26E-2	1,91E-1 ± 6,66E-3	1,91E-1 ± 2,67E-3	1,90E-1 ± 8,29E-3	1,96E-1 ± 4,77E-3
MNSK	2,48E-1 ± 7,33E-2	2,44E-1 ± 3,52E-2	2,33E-1 ± 1,25E-2	2,43E-1 ± 1,12E-2	2,37E-1 ± 9,69E-3	2,44E-1 ± 9,82E-3	2,44E-1 ± 6,81E-3
Lin. Reg.	1,78E-1 ± 2,76E-2	1,83E-1 ± 9,07E-3	1,88E-1 ± 9,55E-3	1,87E-1 ± 2,75E-3	1,90E-1 ± 3,82E-3	1,87E-1 ± 4,83E-3	1,93E-1 ± 3,25E-3
K-NN	2,04E-1 ± 3,76E-2	1,99E-1 ± 1,90E-2	2,00E-1 ± 9,13E-3	2,03E-1 ± 5,98E-3	2,06E-1 ± 3,12E-3	2,02E-1 ± 1,07E-2	2,28E-1 ± 7,93E-3
ZeroR	2,18E-1 ± 3,79E-2	2,19E-1 ± 1,65E-2	2,18E-1 ± 6,88E-3	2,21E-1 ± 5,01E-3	2,18E-1 ± 4,44E-3	2,21E-1 ± 4,80E-3	2,21E-1 ± 2,67E-3
Vstavi 0	4,48E-1 ± 4,90E-2	4,29E-1 ± 2,28E-2	4,23E-1 ± 3,45E-3	4,26E-1 ± 7,53E-3	4,26E-1 ± 1,26E-2	4,27E-1 ± 7,91E-3	4,26E-1 ± 5,13E-3
Bagging	1,74E-1 ± 3,25E-2	1,84E-1 ± 9,65E-3	1,86E-1 ± 1,04E-2	1,88E-1 ± 2,80E-3	1,86E-1 ± 4,46E-3	1,87E-1 ± 5,27E-3	1,95E-1 ± 3,75E-3
Rot. reg. gozd	1,72E-1 ± 3,09E-2	1,80E-1 ± 9,63E-3	1,84E-1 ± 9,86E-3	1,83E-1 ± 3,40E-3	1,83E-1 ± 3,03E-3	1,83E-1 ± 5,95E-3	1,89E-1 ± 4,01E-3
RRG var. 2	1,72E-1 ± 3,01E-2	1,80E-1 ± 9,52E-3	1,84E-1 ± 1,01E-2	1,83E-1 ± 3,43E-3	1,83E-1 ± 3,04E-3	1,83E-1 ± 6,01E-3	1,89E-1 ± 4,18E-3
RRG var. 2	1,75E-1 ± 2,79E-2	1,82E-1 ± 8,05E-3	1,88E-1 ± 1,00E-2	1,86E-1 ± 2,92E-3	1,89E-1 ± 3,97E-3	1,86E-1 ± 6,34E-3	1,97E-1 ± 4,76E-3
SPSS	2,27E-1 ± 7,98E-3	2,51E-1 ± 2,35E-2	2,57E-1 ± 1,28E-2	2,66E-1 ± 3,07E-3	2,74E-1 ± 1,61E-2	2,74E-1 ± 3,31E-3	2,84E-1 ± 7,37E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

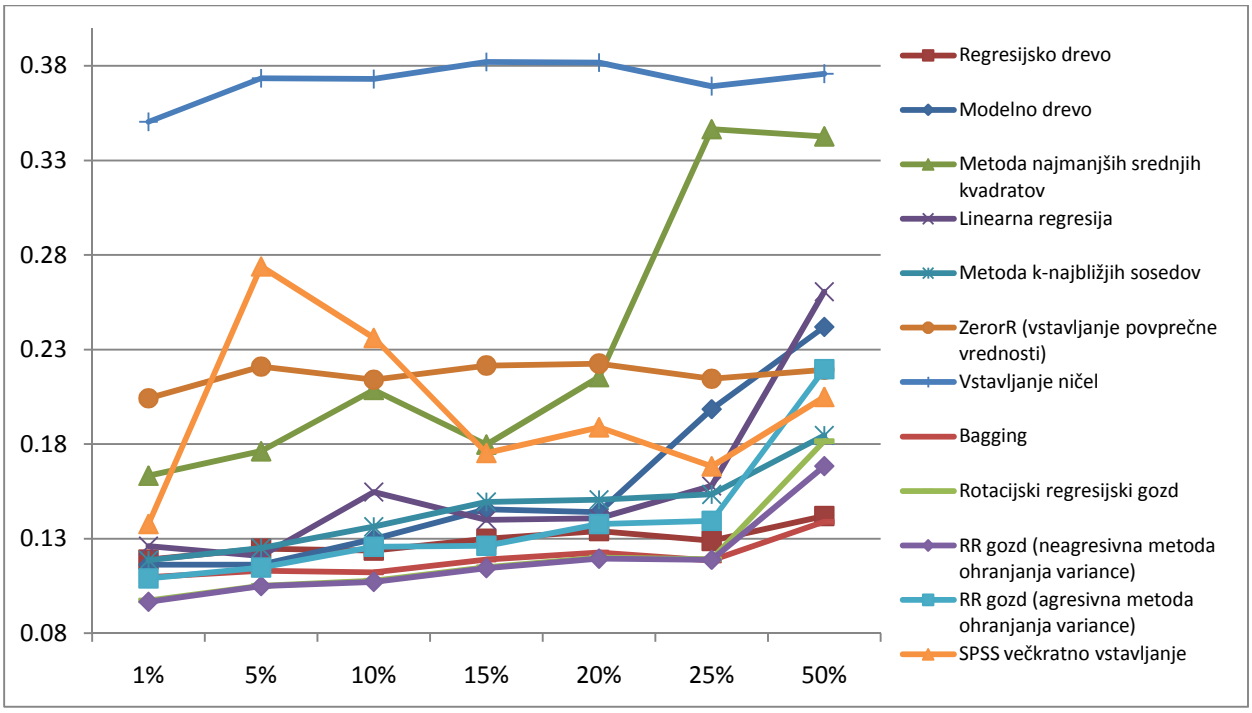


Graf 6.1-10: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Pima Indians (diabetes)* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-11: Povprečne napake po metriki TRMSE na podatkovni množici *Segmentation*

Metoda*	Segment. (1%)	Segment. (5%)	Segment. (10%)	Segment. (15%)	Segment. (20%)	Segment. (25%)	Segment. (50%)
Reg. drevo	1,19E-1 ± 6,00E-2	1,25E-1 ± 3,86E-2	1,24E-1 ± 9,20E-3	1,30E-1 ± 9,92E-3	1,34E-1 ± 8,27E-3	1,29E-1 ± 1,50E-2	1,42E-1 ± 3,72E-3
Modelno drevo	1,16E-1 ± 4,78E-2	1,16E-1 ± 3,73E-2	1,30E-1 ± 2,34E-2	1,46E-1 ± 3,18E-2	1,44E-1 ± 2,23E-2	1,98E-1 ± 1,05E-1	2,42E-1 ± 5,68E-2
MNSK	1,63E-1 ± 8,69E-2	1,76E-1 ± 2,96E-2	2,09E-1 ± 1,53E-1	1,80E-1 ± 3,37E-2	2,15E-1 ± 4,02E-2	3,46E-1 ± 2,35E-1	3,43E-1 ± 1,21E-1
Lin. Reg.	1,26E-1 ± 6,28E-2	1,21E-1 ± 3,87E-2	1,55E-1 ± 3,91E-2	1,40E-1 ± 1,31E-2	1,41E-1 ± 5,79E-3	1,58E-1 ± 2,99E-2	2,61E-1 ± 4,31E-2
K-NN	1,19E-1 ± 3,99E-2	1,25E-1 ± 2,54E-2	1,36E-1 ± 2,37E-2	1,49E-1 ± 9,14E-3	1,51E-1 ± 3,66E-3	1,53E-1 ± 1,63E-2	1,85E-1 ± 1,03E-2
ZeroR	2,04E-1 ± 5,02E-2	2,21E-1 ± 1,54E-2	2,14E-1 ± 1,43E-2	2,22E-1 ± 6,96E-3	2,23E-1 ± 2,73E-3	2,15E-1 ± 1,03E-2	2,19E-1 ± 3,91E-3
Vstavi 0	3,50E-1 ± 5,39E-2	3,73E-1 ± 1,90E-2	3,73E-1 ± 1,12E-2	3,82E-1 ± 4,03E-3	3,82E-1 ± 5,65E-3	3,69E-1 ± 6,89E-3	3,76E-1 ± 6,63E-3
Bagging	1,10E-1 ± 5,35E-2	1,13E-1 ± 3,83E-2	1,12E-1 ± 1,17E-2	1,19E-1 ± 7,39E-3	1,23E-1 ± 4,64E-3	1,19E-1 ± 1,52E-2	1,39E-1 ± 5,41E-3
Rot. reg. gozd	9,73E-2 ± 5,63E-2	1,05E-1 ± 3,19E-2	1,08E-1 ± 1,45E-2	1,15E-1 ± 6,65E-3	1,20E-1 ± 5,04E-3	1,19E-1 ± 1,41E-2	1,82E-1 ± 1,46E-2
RRG var. 2	9,66E-2 ± 5,69E-2	1,05E-1 ± 3,20E-2	1,07E-1 ± 1,39E-2	1,14E-1 ± 6,82E-3	1,19E-1 ± 4,93E-3	1,19E-1 ± 1,37E-2	1,68E-1 ± 8,03E-3
RRG var. 2	1,09E-1 ± 5,29E-2	1,15E-1 ± 3,19E-2	1,26E-1 ± 1,36E-2	1,26E-1 ± 9,82E-3	1,38E-1 ± 8,08E-3	1,39E-1 ± 1,48E-2	2,20E-1 ± 2,52E-2
SPSS	1,38E-1 ± 4,31E-3	2,74E-1 ± 1,20E-1	2,36E-1 ± 8,64E-2	1,75E-1 ± 7,78E-3	1,89E-1 ± 4,67E-3	1,68E-1 ± 1,37E-2	2,05E-1 ± 1,02E-2

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

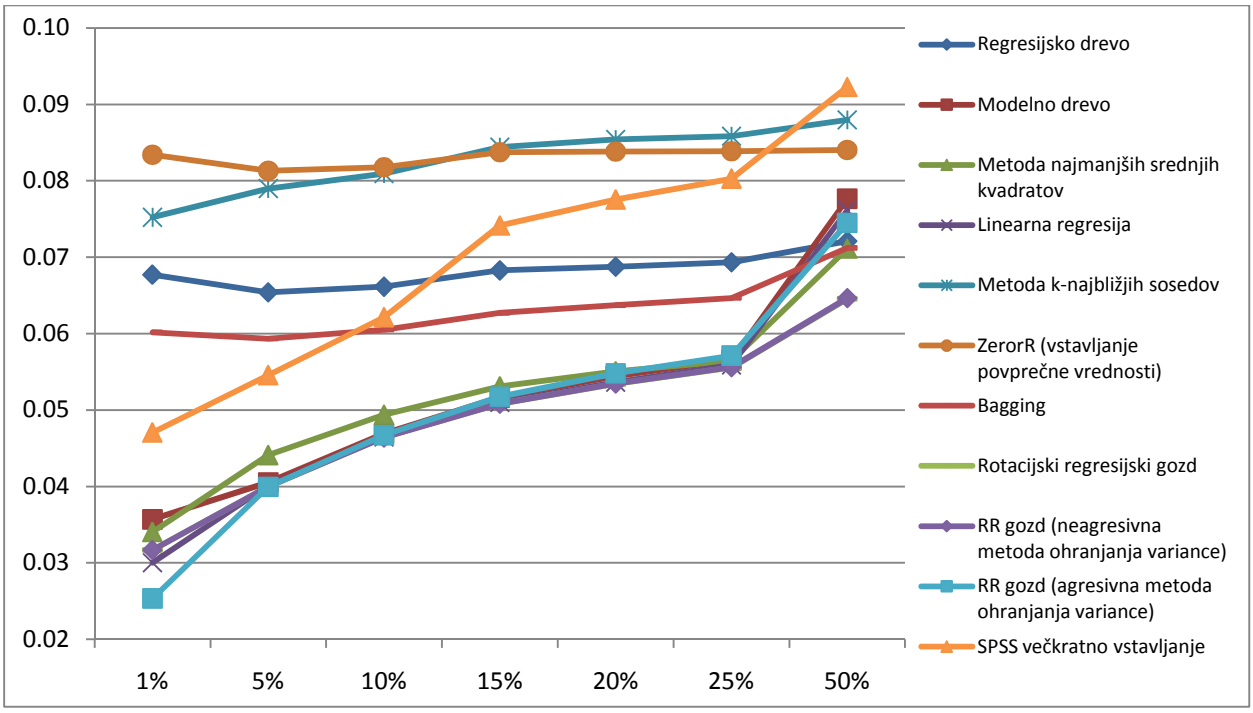


Graf 6.1-11: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Segmentation*

Tabela 6.1-12: Povprečne napake po metriki TRMSE na podatkovni množici *Spellman*

Metoda*	Spellman (1%)	Spellman (5%)	Spellman (10%)	Spellman (15%)	Spellman (20%)	Spellman (25%)	Spellman (50%)
Reg. drevo	6,77E-2 ± 2,79E-3	6,54E-2 ± 4,30E-3	6,61E-2 ± 3,77E-3	6,83E-2 ± 7,35E-4	6,88E-2 ± 1,66E-4	6,93E-2 ± 2,32E-4	7,21E-2 ± 1,07E-3
Modelno drevo	3,57E-2 ± 1,44E-2	4,05E-2 ± 2,17E-3	4,69E-2 ± 2,34E-3	5,16E-2 ± 5,33E-4	5,44E-2 ± 5,40E-4	5,66E-2 ± 2,03E-4	7,76E-2 ± 7,82E-4
MNSK	3,40E-2 ± 1,98E-3	4,41E-2 ± 2,32E-3	4,94E-2 ± 2,34E-3	5,31E-2 ± 6,62E-4	5,51E-2 ± 4,44E-4	5,65E-2 ± 2,96E-4	7,11E-2 ± 5,26E-4
Lin. Reg.	3,00E-2 ± 2,18E-3	4,01E-2 ± 2,01E-3	4,65E-2 ± 2,30E-3	5,10E-2 ± 4,85E-4	5,36E-2 ± 4,45E-4	5,59E-2 ± 3,27E-4	7,62E-2 ± 8,32E-4
K-NN	7,52E-2 ± 1,84E-3	7,90E-2 ± 4,29E-3	8,09E-2 ± 4,42E-3	8,44E-2 ± 1,29E-3	8,54E-2 ± 1,15E-3	8,58E-2 ± 1,57E-3	8,80E-2 ± 1,03E-3
ZeroR	8,34E-2 ± 1,40E-3	8,13E-2 ± 4,45E-3	8,18E-2 ± 3,78E-3	8,37E-2 ± 4,07E-4	8,38E-2 ± 1,97E-4	8,39E-2 ± 2,31E-4	8,40E-2 ± 2,00E-4
Vstavi 0	5,29E-1 ± 5,17E-3	5,31E-1 ± 1,13E-3	5,30E-1 ± 1,73E-3	5,29E-1 ± 3,51E-4	5,30E-1 ± 3,61E-4	5,29E-1 ± 6,87E-4	5,29E-1 ± 2,40E-4
Bagging	6,02E-2 ± 3,00E-3	5,93E-2 ± 3,70E-3	6,05E-2 ± 3,25E-3	6,27E-2 ± 6,59E-4	6,37E-2 ± 4,07E-4	6,47E-2 ± 2,59E-4	7,12E-2 ± 1,89E-4
Rot. reg. gozd	3,17E-2 ± 1,26E-2	4,00E-2 ± 2,15E-3	4,64E-2 ± 2,34E-3	5,08E-2 ± 4,50E-4	5,34E-2 ± 4,71E-4	5,56E-2 ± 3,50E-4	6,46E-2 ± 4,69E-3
RRG var. 2	3,17E-2 ± 1,26E-2	4,00E-2 ± 2,15E-3	4,64E-2 ± 2,34E-3	5,08E-2 ± 4,46E-4	5,35E-2 ± 4,71E-4	5,56E-2 ± 3,47E-4	6,46E-2 ± 4,60E-3
RRG var. 2	2,53E-2 ± 7,53E-4	3,99E-2 ± 2,18E-3	4,67E-2 ± 2,43E-3	5,17E-2 ± 4,02E-4	5,48E-2 ± 4,66E-4	5,71E-2 ± 3,27E-4	7,45E-2 ± 7,18E-4
SPSS	4,71E-2 ± 1,12E-4	5,46E-2 ± 5,22E-4	6,21E-2 ± 3,17E-4	7,41E-2 ± 6,06E-4	7,75E-2 ± 1,74E-4	8,03E-2 ± 3,61E-5	9,23E-2 ± 1,84E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

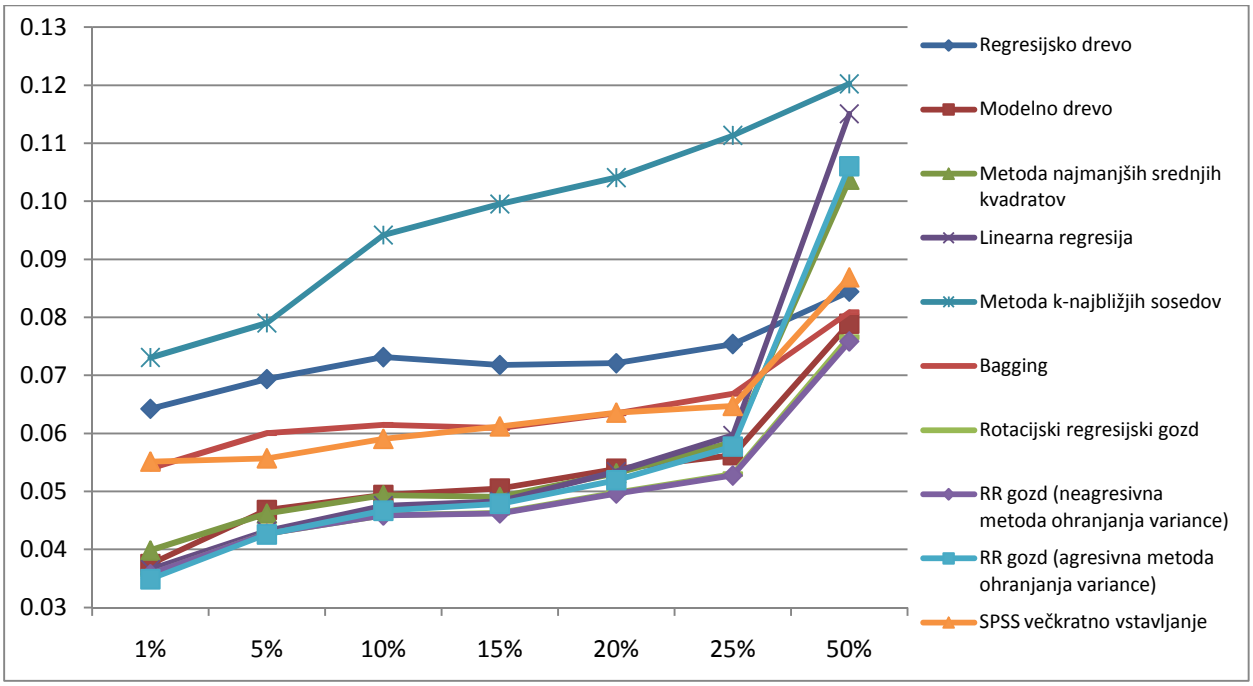


Graf 6.1-12: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Spellman* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.1-13: Povprečne napake po metriki TRMSE na podatkovni množici *Wisconsin breast cancer*

Metoda*	Breast c. (1%)	Breast c. (5%)	Breast c. (10%)	Breast c. (15%)	Breast c. (20%)	Breast c. (25%)	Breast c. (50%)
Reg. drevo	6,42E-2 ± 7,70E-3	6,93E-2 ± 4,04E-3	7,31E-2 ± 2,16E-3	7,18E-2 ± 2,58E-3	7,21E-2 ± 1,90E-3	7,54E-2 ± 1,68E-3	8,44E-2 ± 1,54E-3
Modelno drevo	3,75E-2 ± 6,07E-3	4,68E-2 ± 7,36E-3	4,94E-2 ± 3,58E-3	5,05E-2 ± 1,27E-3	5,39E-2 ± 2,07E-3	5,62E-2 ± 2,00E-3	7,89E-2 ± 1,17E-3
MNSK	3,98E-2 ± 7,64E-3	4,62E-2 ± 6,04E-3	4,94E-2 ± 1,35E-3	4,90E-2 ± 1,42E-3	5,31E-2 ± 2,20E-3	5,85E-2 ± 1,68E-3	1,04E-1 ± 5,19E-3
Lin. Reg.	3,65E-2 ± 7,47E-3	4,31E-2 ± 5,50E-3	4,75E-2 ± 2,66E-3	4,83E-2 ± 2,40E-3	5,34E-2 ± 2,16E-3	5,96E-2 ± 2,04E-3	1,15E-1 ± 4,64E-3
K-NN	7,31E-2 ± 1,48E-2	7,90E-2 ± 5,81E-3	9,42E-2 ± 2,40E-3	9,95E-2 ± 2,40E-3	1,04E-1 ± 2,05E-3	1,11E-1 ± 5,59E-3	1,20E-1 ± 5,44E-3
ZeroR	1,43E-1 ± 1,78E-2	1,41E-1 ± 6,28E-3	1,47E-1 ± 2,15E-3	1,44E-1 ± 3,58E-3	1,44E-1 ± 2,46E-3	1,45E-1 ± 1,82E-3	1,44E-1 ± 1,54E-3
Vstavi 0	2,95E-1 ± 2,07E-2	2,93E-1 ± 6,72E-3	2,94E-1 ± 5,06E-3	2,94E-1 ± 5,86E-3	2,93E-1 ± 3,29E-3	2,97E-1 ± 2,74E-3	2,95E-1 ± 1,82E-3
Bagging	5,39E-2 ± 8,61E-3	6,00E-2 ± 5,19E-3	6,15E-2 ± 1,12E-3	6,09E-2 ± 2,11E-3	6,35E-2 ± 1,80E-3	6,68E-2 ± 2,54E-3	8,09E-2 ± 1,79E-3
Rot. reg. gozd	3,57E-2 ± 7,18E-3	4,27E-2 ± 6,09E-3	4,59E-2 ± 2,63E-3	4,63E-2 ± 1,78E-3	4,98E-2 ± 1,91E-3	5,30E-2 ± 1,69E-3	7,66E-2 ± 1,52E-3
RRG var. 2	3,57E-2 ± 7,11E-3	4,27E-2 ± 6,05E-3	4,59E-2 ± 2,58E-3	4,62E-2 ± 1,75E-3	4,96E-2 ± 1,88E-3	5,27E-2 ± 1,68E-3	7,58E-2 ± 1,52E-3
RRG var. 2	3,49E-2 ± 5,90E-3	4,26E-2 ± 5,28E-3	4,67E-2 ± 1,74E-3	4,79E-2 ± 2,53E-3	5,19E-2 ± 1,77E-3	5,77E-2 ± 2,20E-3	1,06E-1 ± 4,36E-3
SPSS	5,51E-2 ± 1,34E-2	5,57E-2 ± 1,09E-3	5,90E-2 ± 4,03E-4	6,12E-2 ± 3,84E-3	6,36E-2 ± 1,10E-3	6,47E-2 ± 1,77E-3	8,69E-2 ± 1,52E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

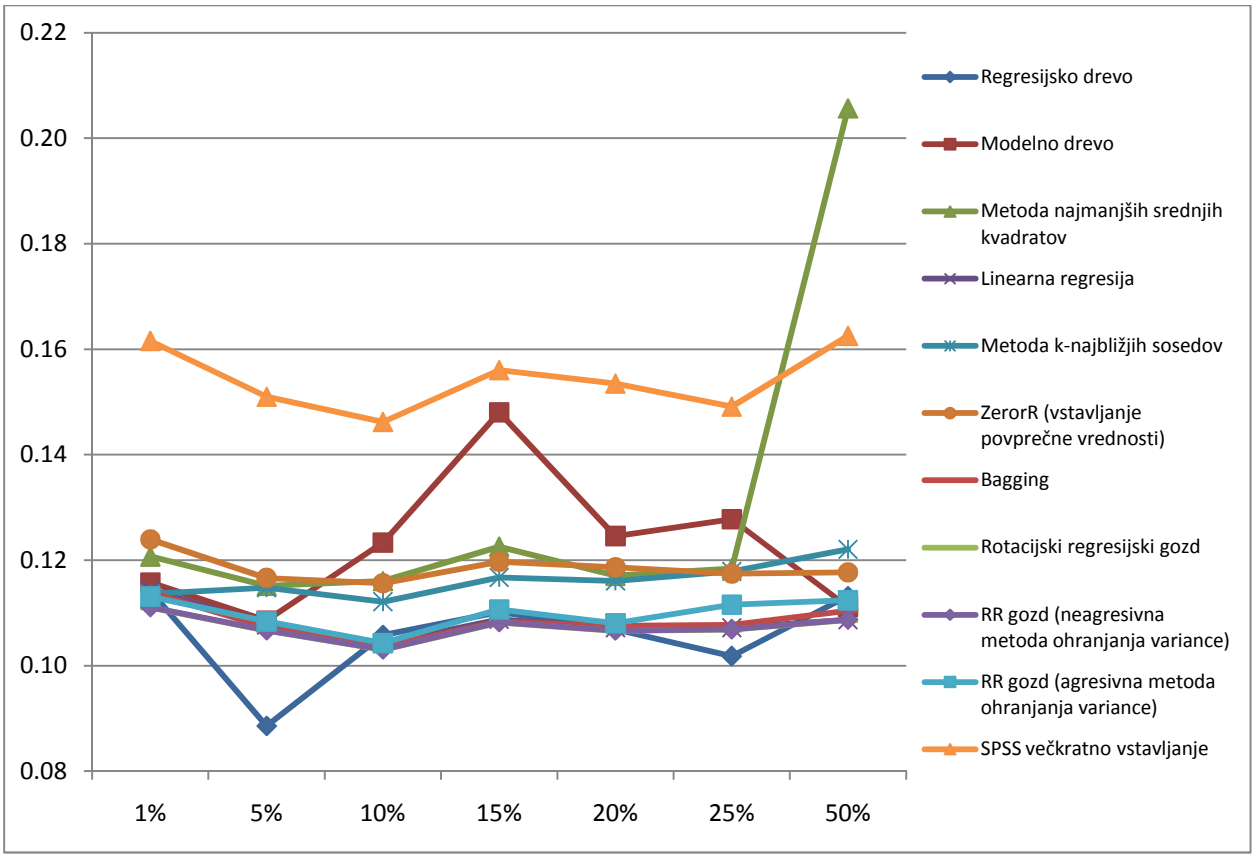


Graf 6.1-13: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Wisconsin breast cancer* (metoda vstavljanja povprečnih vrednosti in metoda vstavljanja ničel zaradi preglednosti grafa nista vključeni, saj njune povprečne napake močno odstopajo)

Tabela 6.1-14: Povprečne napake po metriki TRMSE na podatkovni množici *Yeast*

Metoda*	Yeast (1%)	Yeast (5%)	Yeast (10%)	Yeast (15%)	Yeast (20%)	Yeast (25%)	Yeast (50%)
Reg. drevo	1,20E-1 ± 1,57E-2	1,10E-1 ± 1,54E-2	1,10E-1 ± 4,97E-3	1,12E-1 ± 1,54E-3	1,09E-1 ± 2,61E-3	1,10E-1 ± 5,43E-3	1,12E-1 ± 1,85E-3
Modelno drevo	1,16E-1 ± 1,94E-2	1,09E-1 ± 1,63E-2	1,23E-1 ± 3,79E-2	1,48E-1 ± 5,49E-2	1,25E-1 ± 3,97E-2	1,28E-1 ± 3,27E-2	1,11E-1 ± 5,23E-3
MNSK	1,21E-1 ± 2,20E-2	1,15E-1 ± 1,13E-2	1,16E-1 ± 9,92E-3	1,23E-1 ± 1,17E-2	1,17E-1 ± 4,14E-3	1,18E-1 ± 1,52E-2	2,06E-1 ± 9,47E-2
Lin. Reg.	1,14E-1 ± 2,01E-2	1,08E-1 ± 1,67E-2	1,04E-1 ± 3,80E-3	1,09E-1 ± 1,79E-3	1,07E-1 ± 1,77E-3	1,07E-1 ± 5,23E-3	1,09E-1 ± 2,45E-3
K-NN	1,14E-1 ± 2,26E-2	1,15E-1 ± 1,75E-2	1,12E-1 ± 3,52E-3	1,17E-1 ± 1,38E-3	1,16E-1 ± 9,35E-4	1,18E-1 ± 3,43E-3	1,22E-1 ± 3,77E-3
ZeroR	1,24E-1 ± 1,82E-2	1,17E-1 ± 1,48E-2	1,16E-1 ± 2,93E-3	1,20E-1 ± 1,91E-3	1,19E-1 ± 9,96E-4	1,17E-1 ± 5,66E-3	1,18E-1 ± 1,95E-3
Vstavi 0	3,96E-1 ± 1,51E-2	3,93E-1 ± 8,40E-3	3,93E-1 ± 7,34E-3	3,94E-1 ± 2,81E-3	3,94E-1 ± 4,45E-3	3,92E-1 ± 3,85E-3	3,91E-1 ± 2,19E-3
Bagging	1,14E-1 ± 2,06E-2	1,07E-1 ± 1,64E-2	1,04E-1 ± 3,45E-3	1,08E-1 ± 1,72E-3	1,08E-1 ± 1,32E-3	1,08E-1 ± 5,39E-3	1,10E-1 ± 1,91E-3
Rot. reg. gozd	1,11E-1 ± 2,18E-2	1,07E-1 ± 1,70E-2	1,03E-1 ± 3,77E-3	1,08E-1 ± 1,98E-3	1,07E-1 ± 1,89E-3	1,07E-1 ± 5,15E-3	1,09E-1 ± 2,47E-3
RRG var. 2	1,11E-1 ± 2,20E-2	1,07E-1 ± 1,70E-2	1,03E-1 ± 3,83E-3	1,08E-1 ± 1,99E-3	1,07E-1 ± 1,88E-3	1,07E-1 ± 5,13E-3	1,09E-1 ± 2,45E-3
RRG var. 3	1,13E-1 ± 2,00E-2	1,08E-1 ± 1,72E-2	1,04E-1 ± 4,00E-3	1,11E-1 ± 1,84E-3	1,08E-1 ± 1,76E-3	1,12E-1 ± 6,15E-3	1,12E-1 ± 1,29E-3
SPSS	1,62E-1 ± 1,15E-2	1,51E-1 ± 8,97E-3	1,46E-1 ± 2,61E-4	1,56E-1 ± 1,84E-3	1,53E-1 ± 6,18E-4	1,49E-1 ± 6,42E-4	1,63E-1 ± 8,33E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**



Graf 6.1-14: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Yeast* (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Iz prikazanih podatkov je razvidno, da v večini primerov napaka posameznih metod narašča z naraščanjem odstotka manjkajočih vrednosti. Izjema je predvsem metoda vstavljanja ničel, ki na grafih skorajda ni prikazana zaradi svoje razumljivo visoke povprečne napake. Preprostost metode skupaj z mehanizmom povsem naključnih manjkajočih vrednosti je razlog, da se povprečna napaka te metode bistveno ne spreminja s spreminjanjem odstotka manjkajočih vrednosti. Pričakovano slabo se obnese tudi metoda vstavljanja povprečnih vrednosti, nekoliko bolj presenetljiv pa je slab rezultat algoritma za napovedovanje manjkajočih vrednosti orodja SPSS, ki temelji na gradnji Monte Carlo markovskih verig (*MCMC*). Kot bomo spoznali kasneje, je za to metodo (kakor tudi za druge metode večkratnega vstavljanja) značilno, da se ubada predvsem z ohranjanjem variance (oz. standardne deviacije) in mediane vrednosti posameznih atributov.

Ključnega pomena za naše delo je, kako dobro se je izkazal naš rotacijski regresijski gozd, oziroma njegovi različici z dodatnima metodama za ohranjanje variance. Povprečne napake osnovnega

rotacijskega regresijskega gozda in njegove različice, ki uporablja manj agresivno metodo za ohranjanje variance, se (sodeč po grafih) skoraj ujemajo in so med najnižjimi, če ne celo najnižje. Regresijski gozd z agresivnejšo metodo ohranjanja variance se je odrezal slabše in je po svoji natančnosti kar nekajkrat zaostal za drugimi metodami (predvsem za metodo *bagging*).

Zaradi normalizacije vrednosti atributov podatkovnih množic, ki nam je omogočila uporabo metrike korena povprečne kvadratne napake na celotni množici, so tudi ocene napak majhne, obenem pa je tudi spremenjen vpliv atributov, katerih numerične vrednosti so za razred ali več drugačne od večine preostalih. Za nazornejši prikaz razlik med metodami bi lahko uporabili katero izmed metrik relativnih napak, vendar nanje vpliva variabilnost dejanskih vrednosti, kjer lahko visoka variabilnost vrednosti posameznih atributov tudi privede do navidez majhnih razlik med natančnostmi metod. Mogoče še najbolj očitna primerjava je primerjava metod na podlagi razvrstitve po rangih, ki jih metode dosežejo pri posameznih poskusih. Primer takšne razvrstitve je v tabeli (Tabela 6.1-15), kjer so rangi določeni z neposredno primerjavo povprečnih natančnosti, doseženih pri posameznem poskusu.

Tabela 6.1-15: Razvrstitev metod po rangih glede na natančnost pri posameznih poskusih

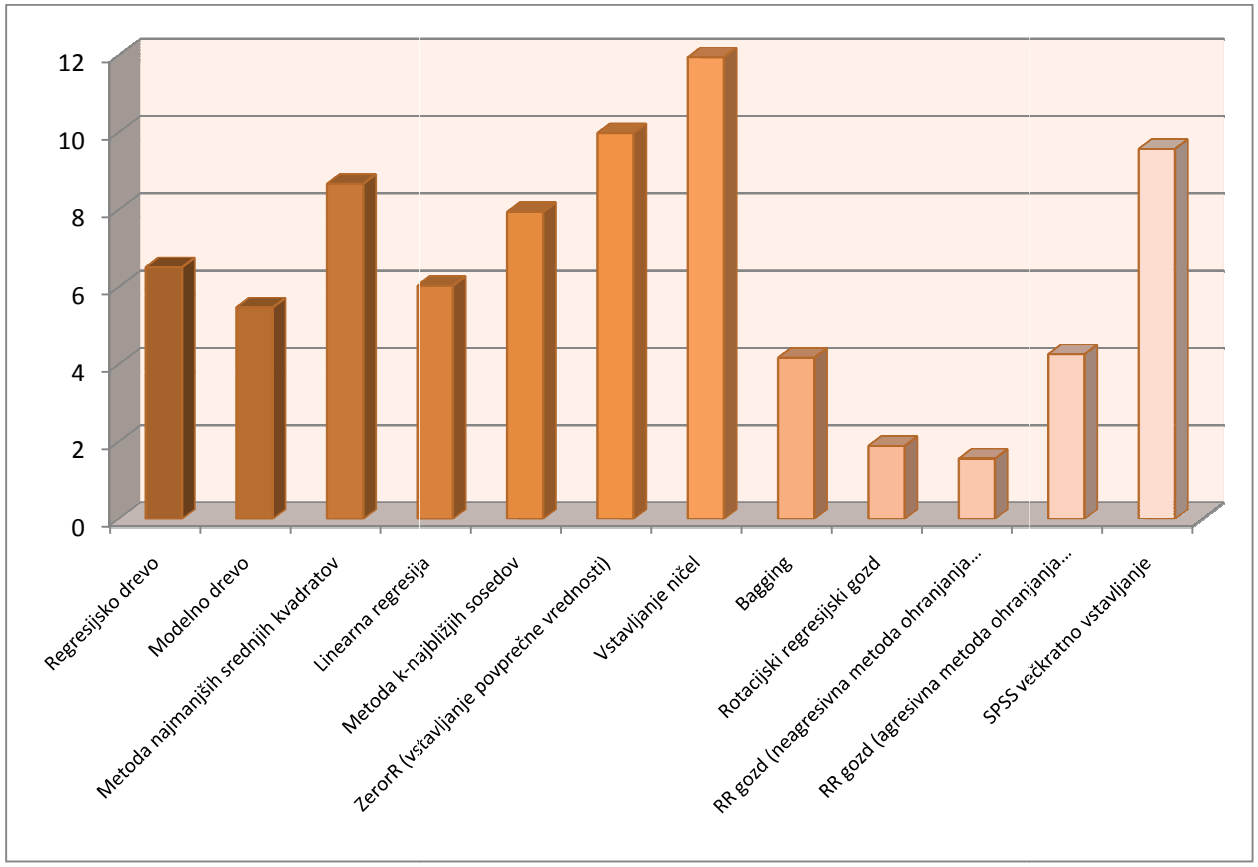
Pod. mn.	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Boh. (1%)	8	7	6	5	9	10	12	4	1	1	3	11
Boh. (5%)	8	6	7	5	9	10	12	3	1	2	4	11
Boh. (10%)	8	7	6	5	10	9	12	3	2	1	4	11
Boh. (15%)	8	7	6	5	10	9	12	3	2	1	4	11
Boh. (20%)	7	8	6	3	10	9	12	4	2	1	5	11
Boh. (25%)	7	8	6	3	10	9	12	4	2	1	5	11
Boh. (50%)	6	8	4	3	10	9	12	5	1	2	7	11
Con. (1%)	8	4	10	7	6	11	12	5	1	2	3	9
Con. (5%)	7	4	10	8	6	11	12	5	2	1	3	9
Con. (10%)	6	4	10	8	7	11	12	5	2	1	3	9
Con. (15%)	5	6	10	7	8	11	12	4	2	1	3	9
Con. (20%)	6	4	10	7	8	11	12	5	2	1	3	9
Con. (25%)	7	4	9	6	8	11	12	5	2	1	3	10
Con. (50%)	6	4	9	5	8	10	12	7	2	1	3	11
Cpu (1%)	5	6	12	8	7	10	11	4	2	1	3	9
Cpu (5%)	7	5	12	8	6	10	11	3	2	1	4	9
Cpu (10%)	5	6	11	7	8	9	12	3	2	1	4	10
Cpu (15%)	6	5	11	7	8	10	12	3	2	1	4	9
Cpu (20%)	6	4	11	7	8	9	12	3	1	2	5	10
Cpu (25%)	7	5	11	6	8	10	12	4	2	1	3	9
Cpu (50%)	3	5	7	8	9	10	12	4	1	2	6	11

Pod. mn.	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
E. coli (1%)	5	8	9	7	6	11	12	4	3	2	1	10
E. coli (5%)	8	6	9	5	7	10	12	3	1	2	4	11
E. coli (10%)	7	9	8	6	5	10	12	4	1	2	3	11
E. coli (15%)	8	5	9	7	6	10	12	3	1	2	4	11
E. coli (20%)	8	7	9	5	6	10	12	3	2	1	4	11
E. coli (25%)	8	4	9	7	6	10	12	3	2	1	5	11
E. coli (50%)	6	7	8	3	9	10	12	5	2	1	4	11
Jap. v. (1%)	7	5	9	8	1	11	12	6	3	2	4	10
Jap. v. (5%)	7	5	9	8	1	11	12	6	3	2	4	10
Jap. v. (10%)	7	5	9	8	3	11	12	6	2	1	4	10
Jap. v. (15%)	7	6	9	8	4	11	12	5	1	2	3	10
Jap. v. (20%)	7	6	9	8	5	11	12	4	1	2	3	10
Jap. v. (25%)	7	4	9	8	6	11	12	5	1	2	3	10
Jap. v. (50%)	8	5	7	6	9	10	12	3	2	1	4	11
Lrs (1%)	7	4	8	6	10	11	12	5	3	1	2	9
Lrs (5%)	7	2	8	6	10	11	12	4	3	1	5	9
Lrs (10%)	7	3	8	6	10	11	12	5	2	1	4	9
Lrs (15%)	7	3	8	6	10	11	12	4	2	1	5	9
Lrs (20%)	7	4	8	6	10	11	12	3	2	1	5	9
Lrs (25%)	6	3	8	7	10	11	12	4	2	1	5	9
Lrs (50%)	2	3	9	10	7	11	12	1	5	4	8	6
M. mas. (1%)	8	5	11	4	6	9	12	7	1	2	3	10
M. mas. (5%)	6	5	11	7	8	9	12	3	2	1	4	10
M. mas. (10%)	7	4	11	6	8	9	12	3	1	2	5	10
M. mas. (15%)	6	4	11	7	8	9	12	3	2	1	5	10
M. mas. (20%)	5	4	10	7	8	9	12	1	2	3	6	11
M. mas. (25%)	4	5	11	8	7	9	12	3	1	2	6	10
M. mas. (50%)	4	5	10	8	6	9	12	3	1	2	7	11
Ozone (1%)	8	4	7	5	10	11	12	6	2	1	3	9
Ozone (5%)	8	3	7	5	10	11	12	6	1	2	4	9
Ozone (10%)	8	4	7	6	10	11	12	5	2	1	3	9
Ozone (15%)	8	4	7	6	10	11	12	3	2	1	5	9
Ozone (20%)	6	4	8	7	10	11	12	2	3	1	5	9
Ozone (25%)	5	2	8	9	10	11	12	1	4	3	6	7
Ozone (50%)	2	3	9	10	7	11	12	1	6	5	8	4
P. bl. (1%)	5	6	12	8	7	10	11	4	2	1	3	9
P. bl. (5%)	6	4	12	8	7	10	11	3	2	1	5	9
P. bl. (10%)	5	7	12	8	6	10	11	3	2	1	4	9
P. bl. (15%)	4	6	11	8	7	9	12	3	2	1	5	10
P. bl. (20%)	4	7	12	8	6	9	11	3	2	1	5	10
P. bl. (25%)	4	5	11	8	7	9	12	3	2	1	6	10
P. bl. (50%)	4	6	11	8	9	7	12	3	2	1	5	10

Pod. mn.	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
P. Ind. (1%)	7	5	11	6	8	9	12	3	1	2	4	10
P. Ind. (5%)	7	6	10	4	8	9	12	5	1	2	3	11
P. Ind. (10%)	7	6	10	5	8	9	12	3	1	2	4	11
P. Ind. (15%)	8	6	10	4	7	9	12	5	2	1	3	11
P. Ind. (20%)	7	6	10	5	8	9	12	3	2	1	4	11
P. Ind. (25%)	7	6	10	5	8	9	12	4	1	2	3	11
P. Ind. (50%)	7	5	10	3	9	8	12	4	1	2	6	11
Segm. (1%)	7	5	10	8	6	11	12	4	2	1	3	9
Segm. (5%)	7	5	9	6	8	10	12	3	2	1	4	11
Segm. (10%)	4	6	9	8	7	10	12	3	2	1	5	11
Segm. (15%)	5	7	10	6	8	11	12	3	2	1	4	9
Segm. (20%)	4	7	10	6	8	11	12	3	2	1	5	9
Segm. (25%)	4	9	11	7	6	10	12	1	3	2	5	8
Segm. (50%)	2	9	11	10	5	7	12	1	4	3	8	6
Spell. (1%)	9	6	5	2	10	11	12	8	4	3	1	7
Spell. (5%)	9	5	6	4	10	11	12	8	3	2	1	7
Spell. (10%)	9	5	6	3	10	11	12	7	2	1	4	8
Spell. (15%)	8	4	6	3	11	10	12	7	2	1	5	9
Spell. (20%)	8	4	6	3	11	10	12	7	1	2	5	9
Spell. (25%)	8	5	4	3	11	10	12	7	1	2	6	9
Spell. (50%)	5	8	3	7	10	9	12	4	1	2	6	11
Wisc. (1%)	9	5	6	4	10	11	12	7	2	3	1	8
Wisc. (5%)	9	6	5	4	10	11	12	8	2	3	1	7
Wisc. (10%)	9	6	5	4	10	11	12	8	2	1	3	7
Wisc. (15%)	9	6	5	4	10	11	12	7	2	1	3	8
Wisc. (20%)	9	6	4	5	10	11	12	7	2	1	3	8
Wisc. (25%)	9	3	5	6	10	11	12	8	2	1	4	7
Wisc. (50%)	5	3	7	9	10	11	12	4	2	1	8	6
Yeast (1%)	8	7	9	6	4	10	12	5	1	2	3	11
Yeast (5%)	7	6	9	4	8	10	12	3	1	2	5	11
Yeast (10%)	6	10	9	4	7	8	12	3	1	2	5	11
Yeast (15%)	6	10	9	4	7	8	12	3	1	2	5	11
Yeast (20%)	6	10	8	3	7	9	12	4	1	2	5	11
Yeast (25%)	5	10	9	3	8	7	12	4	1	2	6	11
Yeast (50%)	6	5	11	3	9	8	12	4	1	2	7	10
Povprečni rang	**6,51**	**5,47**	**8,67**	**6,03**	**7,94**	**9,98**	**11,94**	**4,17**	**1,89**	**1,56**	**4,26**	**9,57**

Iz tabele lahko razberemo, da je bila v kar 90 od 98 poskusov najbolje rangirana metoda ena izmed različic rotacijskega regresijskega gozda. Edini preostali metodi, ki sta občasno prekašali rotacijski

regresijski gozd, sta bili *bagging* (6 primerov) in metoda k-najbližjih sosedov (2 primera), ki je svoje najboljše rezultate zabeležila na eni podatkovni množici (*japanese vowels*), medtem ko je po svojem povprečnem rangu v spodnji polovici primerjanih metod (Graf 6.1-15).



Graf 6.1-15: Povprečni rangi metod za napovedovanje manjkajočih vrednosti glede na njihovo natančnost

Za statistično zavrnitev hipoteze, da so posamezne metode med seboj enakovredne, lahko uporabimo neparametrični Friedmanov test, ki smo ga opisali v 2. poglavju. Ker je število poskusov kot tudi metod dovolj veliko, lahko aproksimiramo verjetnostno porazdelitev testne statistike, ki jo izračunamo po formuli 2.20 s porazdelitvijo χ^2 in določimo vrednost *p* (Tabela 6.1-16).

Tabela 6.1-16: Friedmanov test za primerjavo metod po natančnosti napovedovanja manjkajočih vrednosti

N	χ^2	stopnje prostosti	p
98	868,593	11	0,000

Če sedaj želimo dokazati statistično signifikantne razlike med posameznimi metodami, jih lahko paroma primerjamo med sabo s pomočjo Wilcoxonovega testa predznačenih rangov (znan tudi kot Wilcoxonov test enakovrednih parov). Tudi ta neparametrični test je bil predstavljen v drugem poglavju. Ker je v našem primeru število parov dovolj veliko (> 25), lahko s pomočjo formule 2.17 preverimo ničelno hipotezo, da se izbrani metodi v svoji natančnosti bistveno ne razlikujeta. Na podlagi povprečnih rangov smo izbrali 3 metode, ki so po svoji natančnosti najbliže rotacijskemu regresijskemu gozdu: bagging, modelno drevo in linearna regresija. Vsako izmed teh metod smo s pomočjo Wilcoxonovega testa primerjali z vsako izmed naših treh variant rotacijskega regresijskega gozda. Rezultati analize, opravljene z orodjem za statistično analizo SPSS, so predstavljeni v tabeli.

Tabela 6.1-17: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo rotacijskega regresijskega gozda s preostalimi najboljšimi metodami (N=98)

1. metoda	2. metoda	Z	Signifikanca
Linearna regresija	Rotacijski regresijski gozd	-8,524[a]	0,000
Modelno drevo	Rotacijski regresijski gozd	-8,024[a]	0,000
Bagging	Rotacijski regresijski gozd	-7,457[a]	0,000
Linearna regresija	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-8,524[a]	0,000
Modelno drevo	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-8,046[a]	0,000
Bagging	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-7,489[a]	0,000
Linearna regresija	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-7,287[a]	0,000
Modelno drevo	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-4,548[a]	0,000
Bagging	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-0,019[b]	0,984
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-3,395[b]	0,001
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-8,134[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-8,056[a]	0,000

[a] na podlagi negativnih rangov
[b] na podlagi pozitivnih rangov

Ob upoštevanju pogoja $z < -1{,}96$ pri $p < 0{,}05$ je razvidno, da se tako osnovni rotacijski regresijski gozd kot njegova različica z neagresivno metodo za ohranjanje variance signifikantno razlikujeta od preostalih treh metod. Tudi varianta rotacijskega regresijskega gozda z agresivno metodo za ohranjanje variance, ki je po svoji natančnosti najšibkejša izmed treh, se na podlagi Wilcoxonovega testa še vedno signifikantno razlikuje od linearne regresije in modelnega drevesa, ter je po svoji sposobnosti numerične

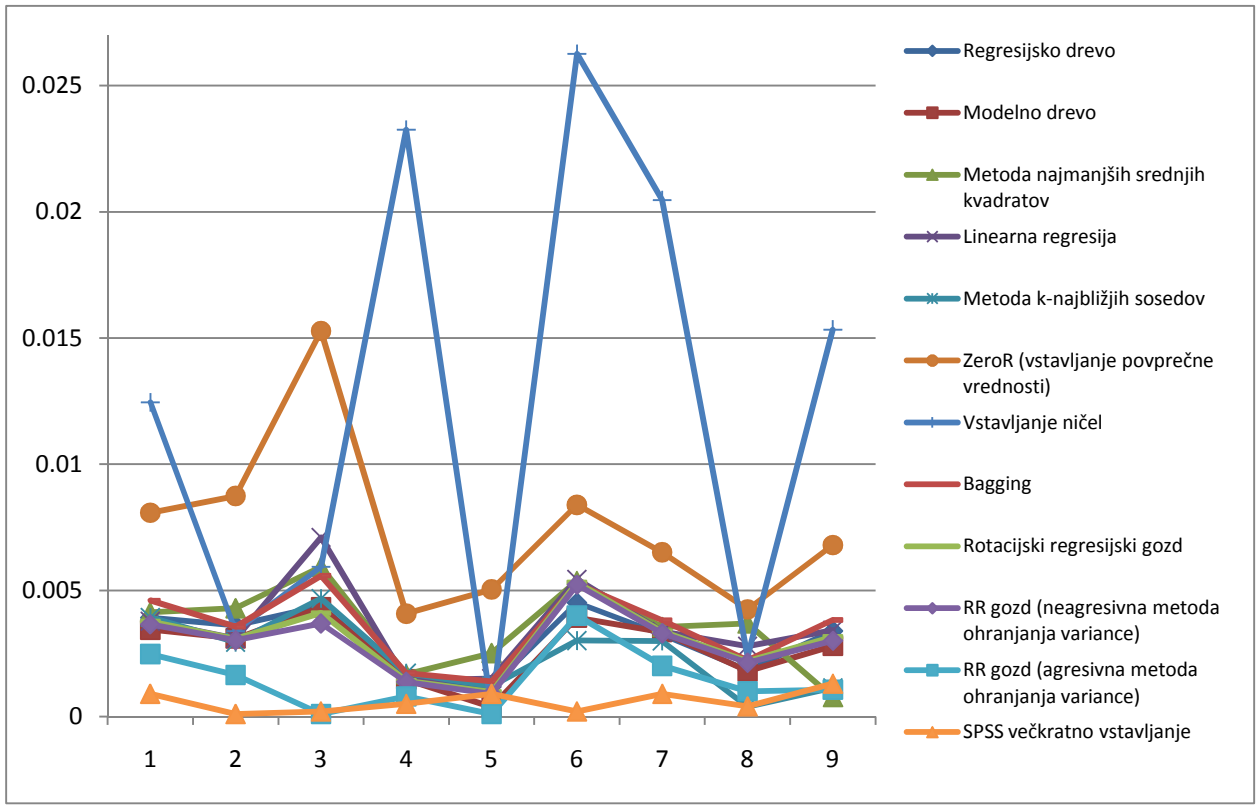
predikcije primerljiva z metodo bagging. Iz primerjave variant rotacijskega regresijskega gozda je razvidno, da se vse tri med seboj signifikantno razlikujejo, pri čemer se je kot najnatančnejša izkazala različica z neagresivno metodo za ohranjanje variance.

6.1.1 Ohranjanje variance

V tretjem poglavju smo predstavili pomen ohranjanja variance pri nadomeščanju manjkajočih vrednosti v podatkovnih množicah. Popolna metoda za nadomeščanje manjkajočih vrednosti bi pravilno napovedala vse vrednosti, torej bi bila njena skupna napaka enaka 0, pri čemer bi ohranila nespremenjene variance vrednosti vseh atributov. Ohranjanje variance smo ocenjevali s primerjanjem (računanjem razlike) varianc vrednosti posameznih atributov v podatkovni množici pred nadomeščanjem manjkajočih vrednosti in po njem. Ker je skupno število vseh atributov v izbranih podatkovnih množicah razmeroma veliko (več kot 400), je skupno število vseh primerjanj preseglo 160.000. Računanje skupne variance za posamezno podatkovno množico zaradi raznolikosti atributov ni možno, zato smo za vsak atribut pri vsakem poskusu razvrstili metode glede na njihovo uspešnost ohranjanja variance in nato izračunali povprečen rang posamezne metode za celoten poskus. Tako smo skupno število primerjav metod »zmanjšali« na obvladljivih 490 primerjav po rangih. Primer izračuna rangov iz razlik varianc je prikazan na grafu (Graf 6.1-16) in v tabeli (Tabela 6.1-18). Na prikazanem primeru se je po stopnji ohranjanja variance najbolje izkazala metoda programskega paketa SPSSS, ki ji sledi različica rotacijskega regresijskega gozda z agresivnim pristopom ohranjanja variance. Večina preostalih metod je med seboj primerljivih, medtem ko sta se pričakovano najslabše odrezali metodi, ki vstavljata povprečne vrednosti oziroma ničle. Podobno bi lahko razbrali iz večine izmed preostalih 489 grafov, ki bi ponazarjali ohranjanje variance pri posameznem poskusu. Seveda jih vseh ne moremo prikazati, vendar preden se lotimo analize na podlagi povprečnih rangov, si na primeru oglejmo še, kako na ohranjanje variance pri posameznih metodah vpliva delež manjkajočih vrednosti.

Iz grafa (Graf 6.1-17) je razvidno, da se pri izbranem atributu stopnja ohranjanja variance slabša z večanjem deleža manjkajočih vrednosti. Spet je najboljša metoda programskega paketa SPSS, s katero se lahko primerja kvečjemu različica rotacijskega regresijskega gozda z agresivnim pristopom ohranjanja variance. Pri manjših deležih manjkajočih vrednosti (do 10% ali 15%) se skoraj vse metode obnesejo precej dobro, medtem ko so pri višjih deležih razlike že opazne, še posebej pri 50% manjkajočih vrednosti. Pričakovano najslabši sta zopet najbolj preprosti metodi, ki vstavljata povprečne vrednosti oziroma ničle. Pri prvi lahko pri vsakem poskusu pričakujemo občutno zmanjšanje variance vrednosti

izbranega atributa kot posledico vstavljanja povprečne vrednosti, medtem ko je pri drugi vpliv odvisen od začetne porazdelitve vrednosti. Tako je možno, da pri samih variancah pred in po nadomeščanju manjkajočih vrednosti ni opaziti občutne spremembe, vendar ima vstavljanje ničel očiten vpliv na skupno povprečno vrednost posameznega atributa, kar lahko razberemo iz grafa (Graf 6.1-18). Na konkretnem primeru so preostale metode dobro ohranile začetno povprečno vrednost izbranega atributa.



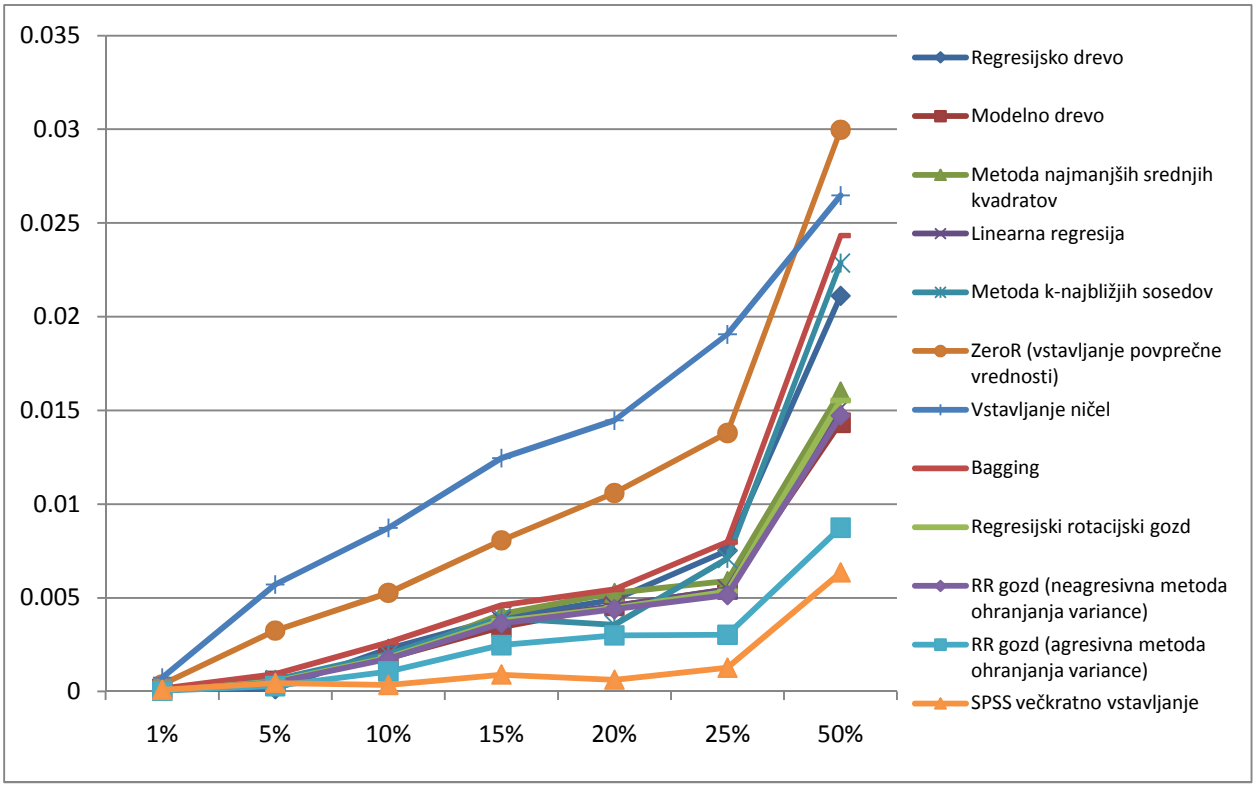
Graf 6.1-16: Primerjava ohranjanja variance (absolutna razlika variance pred in po nadomeščanju) na podatkovni množici *Concrete* (9 numeričnih atributov) pri 15% manjkajočih vrednosti

Končno oceno vsake izmed metod na posamezni množici pri posamezni stopnji manjkajočih vrednosti smo izračunali kot aritmetično sredino povprečnih rangov po Friedmanovem testu, dobljenih na podlagi petih nadomeščanj. Tako dobljene ocene se nahajajo v tabeli (Tabela 6.1-19).

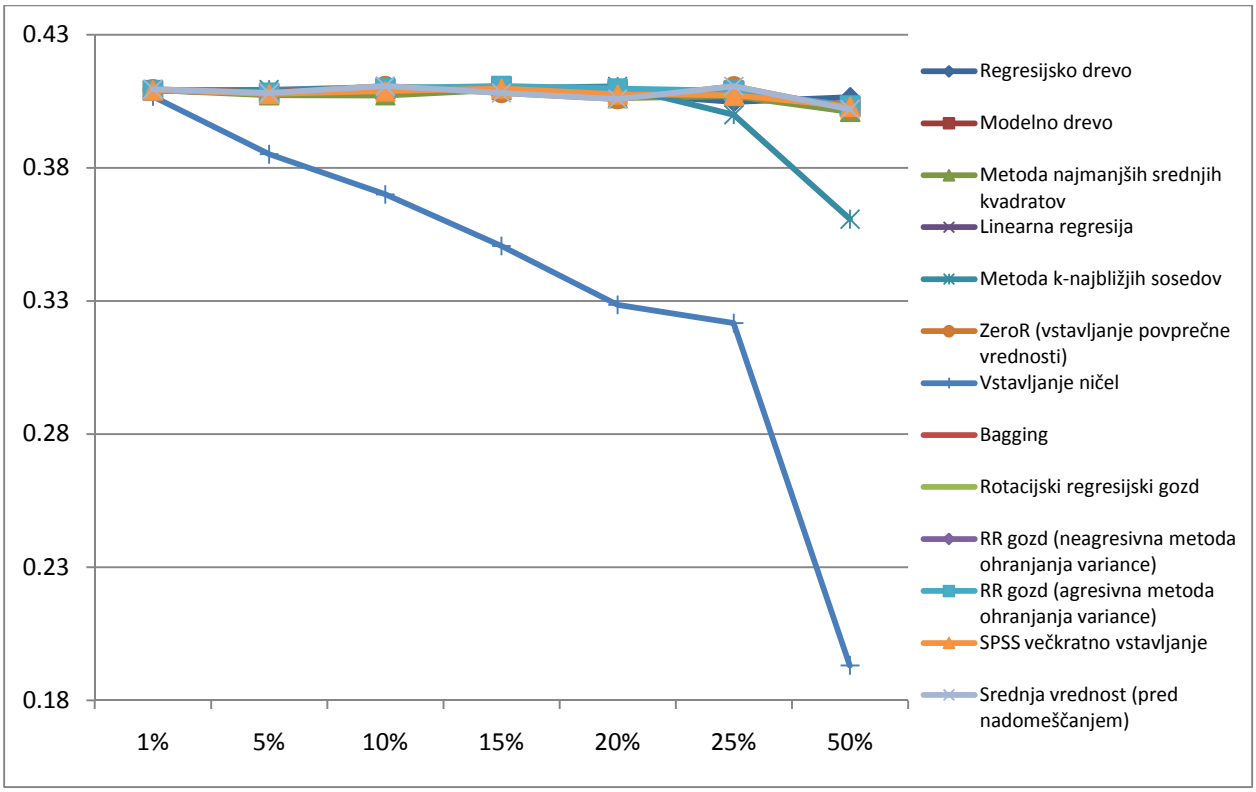
Tabela 6.1-18: Absolutne razlike varianc pred in po nadomeščanju manjkajočih vrednosti ter povprečni rangi metod na podatkovni množici *Concrete* (9 numeričnih atributov) pri 15% manjkajočih vrednosti

Atr.	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
1	3,92E-3 ± 2,78E-4	3,44E-3 ± 1,75E-4	4,13E-3 ± 3,03E-4	3,77E-3 ± 6,50E-5	3,93E-3 ± 1,27E-4	8,07E-3 ± 8,13E-4	1,25E-2 ± 9,01E-4	4,59E-3 ± 2,83E-4	3,76E-3 ± 6,80E-5	3,65E-3 ± 1,02E-4	2,48E-3 ± 2,22E-4	6,11E-3 ± 4,24E-3
2	3,61E-3 ± 1,07E-3	3,10E-3 ± 1,36E-3	4,29E-3 ± 2,62E-4	3,10E-3 ± 2,71E-4	2,95E-3 ± 5,01E-4	8,74E-3 ± 5,91E-4	3,39E-3 ± 4,50E-4	3,56E-3 ± 3,97E-4	3,08E-3 ± 7,30E-4	3,02E-3 ± 9,18E-4	1,65E-3 ± 6,25E-4	9,46E-5 ± 5,35E-4
3	4,37E-3 ± 6,05E-4	4,33E-3 ± 7,46E-4	5,94E-3 ± 6,98E-4	7,10E-3 ± 9,36E-4	4,67E-3 ± 5,33E-4	1,53E-2 ± 1,54E-3	5,94E-3 ± 6,98E-4	5,58E-3 ± 2,17E-4	4,10E-3 ± 3,55E-4	3,69E-3 ± 3,27E-4	9,53E-5 ± 6,51E-4	1,75E-3 ± 6,92E-4
4	1,47E-3 ± 1,68E-5	1,47E-3 ± 1,30E-4	1,67E-3 ± 7,37E-5	1,43E-3 ± 4,10E-5	1,72E-3 ± 3,49E-4	4,08E-3 ± 2,10E-4	2,33E-2 ± 6,65E-4	1,73E-3 ± 1,28E-4	1,42E-3 ± 7,12E-5	1,37E-3 ± 7,95E-5	7,97E-4 ± 1,11E-5	3,54E-3 ± 2,32E-3
5	1,19E-3 ± 3,92E-4	3,72E-4 ± 6,96E-4	2,51E-3 ± 4,63E-4	1,51E-3 ± 1,25E-4	1,21E-3 ± 1,72E-5	5,04E-3 ± 2,69E-4	3,16E-4 ± 5,54E-5	1,39E-3 ± 4,12E-4	9,74E-4 ± 3,07E-4	8,94E-4 ± 3,13E-4	9,88E-5 ± 2,08E-5	6,46E-3 ± 4,32E-3
6	4,51E-3 ± 1,04E-3	3,92E-3 ± 1,89E-2	5,38E-3 ± 8,94E-4	5,44E-3 ± 4,56E-4	3,01E-3 ± 6,79E-4	8,39E-3 ± 7,82E-4	2,63E-2 ± 2,65E-3	5,26E-3 ± 4,26E-4	5,31E-3 ± 7,78E-4	5,23E-3 ± 8,20E-4	4,00E-3 ± 3,29E-4	1,22E-3 ± 5,34E-4
7	3,24E-3 ± 1,35E-3	3,35E-3 ± 9,21E-4	3,55E-3 ± 2,56E-4	3,37E-3 ± 1,94E-4	2,99E-3 ± 6,97E-5	6,51E-3 ± 2,18E-4	2,05E-2 ± 2,26E-4	3,82E-3 ± 7,57E-4	3,37E-3 ± 8,30E-4	3,30E-3 ± 9,26E-4	2,01E-3 ± 1,38E-5	5,45E-3 ± 3,68E-3
8	1,84E-3 ± 1,12E-3	1,79E-3 ± 6,58E-4	3,69E-3 ± 1,61E-4	2,79E-3 ± 9,15E-5	3,94E-4 ± 1,59E-4	4,25E-3 ± 2,40E-4	2,39E-3 ± 1,40E-4	2,23E-3 ± 4,33E-4	2,19E-3 ± 3,24E-4	2,13E-3 ± 3,37E-4	1,00E-3 ± 1,67E-5	2,78E-3 ± 1,62E-3
9	3,39E-3 ± 3,97E-4	2,80E-3 ± 2,71E-4	7,70E-4 ± 4,94E-4	3,42E-3 ± 1,87E-4	1,14E-3 ± 1,05E-3	6,79E-3 ± 1,46E-4	1,53E-2 ± 6,44E-4	3,83E-3 ± 1,03E-5	3,15E-3 ± 2,20E-5	3,00E-3 ± 1,36E-5	1,06E-3 ± 2,51E-4	9,07E-3 ± 5,96E-3
Rang*	6,44	4,67	8,67	8,33	5,00	11,44	9,78	9,00	6,00	4,56	2,11	1,89

*) Povprečni rang po Friedmanovem testu (N=9, χ^2=69,199, sp=11, sig.=0,000)



Graf 6.1-17: Primerjava ohranjanja variance glede na delež manjkajočih vrednosti 1. numeričnega atributa podatkovne množice *Concrete*.



Graf 6.1-18: Vpliv deleža manjkajočih vrednosti na ohranjanje srednje vrednosti 1. atributa podatkovne množice *Concrete*.

Tabela 6.1-19: Razvrstitev metod po rangih (povprečje 5 meritev) glede na uspešnost ohranjanja variance

Po. m.	RD	MD	LMS	LR	KNN	Pov.	ZeroR	Bag.	RRG	RRG1	RRG2	SPSS
Boh. (1%)	5,19E+0 ± 2,56E+0	2,59E+0 ± 3,09E-1	9,25E+0 ± 1,77E-1	6,78E+0 ± 3,09E-1	3,78E+0 ± 9,28E-1	1,10E+1 ± 4,42E-2	1,20E+1 ± 0,00E+0	9,13E+0 ± 4,42E-1	5,19E+0 ± 6,19E-1	5,19E+0 ± 6,19E-1	4,53E+0 ± 3,98E-1	1,53E+0 ± 1,33E-1
Boh. (5%)	8,09E+0 ± 4,42E-2	3,41E+0 ± 1,33E-1	8,78E+0 ± 3,09E-1	7,06E+0 ± 8,84E-2	3,47E+0 ± 3,09E-1	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,50E+0 ± 1,77E-1	5,81E+0 ± 8,84E-2	4,78E+0 ± 4,42E-2	2,66E+0 ± 4,42E-2	1,00E+0 ± 0,00E+0
Boh. (10%)	8,25E+0 ± 3,54E-1	3,53E+0 ± 1,33E-1	8,31E+0 ± 1,77E-1	6,84E+0 ± 1,33E-1	4,19E+0 ± 3,54E-1	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,84E+0 ± 4,42E-2	6,03E+0 ± 4,42E-2	4,75E+0 ± 0,00E+0	2,19E+0 ± 0,00E+0	1,00E+0 ± 0,00E+0
Boh. (15%)	8,69E+0 ± 3,54E-1	3,31E+0 ± 1,77E-1	8,16E+0 ± 4,42E-2	6,69E+0 ± 0,00E+0	5,09E+0 ± 4,86E-1	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,78E+0 ± 4,42E-2	5,72E+0 ± 1,33E-1	4,53E+0 ± 1,33E-1	2,03E+0 ± 4,42E-2	1,00E+0 ± 0,00E+0
Boh. (20%)	8,13E+0 ± 1,33E+0	2,91E+0 ± 3,98E-1	8,13E+0 ± 7,07E-1	6,44E+0 ± 8,84E-2	5,88E+0 ± 8,84E-2	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,75E+0 ± 8,84E-2	5,94E+0 ± 4,42E-1	4,66E+0 ± 2,21E-1	2,19E+0 ± 2,65E-1	1,00E+0 ± 0,00E+0
Boh. (25%)	8,47E+0 ± 7,51E-1	2,81E+0 ± 3,54E-1	7,66E+0 ± 3,98E-1	6,16E+0 ± 3,09E-1	6,78E+0 ± 3,98E-1	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,78E+0 ± 1,33E-1	5,63E+0 ± 8,84E-2	4,47E+0 ± 4,42E-2	2,25E+0 ± 3,54E-1	1,00E+0 ± 0,00E+0
Boh. (50%)	8,91E+0 ± 1,33E-1	3,59E+0 ± 1,10E+0	6,97E+0 ± 4,42E-2	5,13E+0 ± 3,54E-1	7,53E+0 ± 2,21E-1	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,78E+0 ± 4,42E-2	5,72E+0 ± 4,86E-1	4,25E+0 ± 7,07E-1	2,13E+0 ± 1,77E-1	1,00E+0 ± 0,00E+0
Con. (1%)	5,89E+0 ± 1,41E+0	4,78E+0 ± 1,10E+0	6,83E+0 ± 7,86E-2	5,33E+0 ± 1,41E+0	7,06E+0 ± 1,96E+0	1,09E+1 ± 3,93E-1	8,89E+0 ± 1,10E+0	6,06E+0 ± 1,18E+0	5,28E+0 ± 8,64E-1	4,78E+0 ± 6,29E-1	3,39E+0 ± 7,86E-2	7,28E+0 ± 3,93E-1
Con. (5%)	4,28E+0 ± 3,93E-1	4,67E+0 ± 1,26E+0	9,78E+0 ± 1,57E-1	8,72E+0 ± 2,36E-1	4,83E+0 ± 7,86E-2	1,14E+1 ± 0,00E+0	9,44E+0 ± 3,14E-1	6,67E+0 ± 1,57E-1	6,78E+0 ± 3,14E-1	5,33E+0 ± 4,71E-1	2,50E+0 ± 3,93E-1	3,28E+0 ± 2,28E+0
Con. (10%)	6,89E+0 ± 6,29E-1	5,11E+0 ± 7,86E-1	9,33E+0 ± 0,00E+0	7,56E+0 ± 3,14E-1	5,61E+0 ± 3,93E-1	1,14E+1 ± 0,00E+0	9,44E+0 ± 4,71E-1	8,17E+0 ± 3,93E-1	6,11E+0 ± 1,57E-1	4,67E+0 ± 0,00E+0	1,61E+0 ± 2,36E-1	1,83E+0 ± 2,36E-1
Con. (15%)	5,61E+0 ± 1,18E+0	4,78E+0 ± 1,57E-1	8,56E+0 ± 1,57E-1	8,44E+0 ± 1,57E-1	5,33E+0 ± 4,71E-1	1,14E+1 ± 7,86E-2	9,78E+0 ± 0,00E+0	9,06E+0 ± 7,86E-2	6,28E+0 ± 3,93E-1	4,72E+0 ± 2,36E-1	2,22E+0 ± 1,57E-1	1,67E+0 ± 3,14E-1
Con. (20%)	6,78E+0 ± 4,71E-1	5,28E+0 ± 7,86E-2	7,78E+0 ± 1,57E-1	8,06E+0 ± 2,36E-1	6,06E+0 ± 7,07E-1	1,14E+1 ± 0,00E+0	9,00E+0 ± 0,00E+0	8,72E+0 ± 2,36E-1	6,11E+0 ± 3,14E-1	4,83E+0 ± 2,36E-1	2,11E+0 ± 4,71E-1	1,33E+0 ± 3,14E-1
Con. (25%)	7,50E+0 ± 7,86E-2	4,83E+0 ± 2,36E-1	7,06E+0 ± 2,36E-1	7,56E+0 ± 7,86E-1	7,44E+0 ± 1,73E+0	1,14E+1 ± 0,00E+0	8,67E+0 ± 4,71E-1	8,50E+0 ± 3,93E-1	6,33E+0 ± 4,71E-1	5,06E+0 ± 5,50E-1	2,06E+0 ± 7,86E-2	1,28E+0 ± 7,86E-2
Con.	7,78E+0 ±	3,50E+0 ±	6,56E+0 ±	5,94E+0 ±	8,56E+0 ±	1,16E+1 ±	8,94E+0 ±	1,02E+1 ±	6,28E+0 ±	4,78E+0 ±	2,44E+0 ±	1,28E+0 ±

Po. m.	RD	MD	LMS	LR	KNN	Pov.	ZeroR	Bag.	RRG	RRG1	RRG2	SPSS
(50%)	0,00E+0	7,07E-1	7,86E-1	3,93E-1	6,29E-1	1,57E-1	2,36E-1	2,36E-1	3,93E-1	1,57E-1	1,57E-1	3,93E-1
Cpu (1%)	6,23E+0 ± 8,42E-2	6,36E+0 ± 2,32E+0	7,06E+0 ± 3,20E-1	6,75E+0 ± 1,06E+0	6,81E+0 ± 2,09E+0	1,12E+1 ± 7,41E-1	2,57E+0 ± 2,22E+0	7,44E+0 ± 1,03E+0	6,21E+0 ± 1,11E+0	5,35E+0 ± 6,90E-1	3,56E+0 ± 3,87E-1	7,21E+0 ± 3,03E-1
Cpu (5%)	5,00E+0 ± 2,02E-1	6,43E+0 ± 1,62E+0	9,43E+0 ± 8,08E-1	6,43E+0 ± 6,06E-1	6,14E+0 ± 2,02E-1	1,13E+1 ± 6,06E-1	5,36E+0 ± 3,03E-1	6,50E+0 ± 1,11E+0	6,79E+0 ± 7,07E-1	5,64E+0 ± 3,03E-1	3,29E+0 ± 4,04E-1	5,43E+0 ± 0,00E+0
Cpu (10%)	6,21E+0 ± 1,72E+0	7,07E+0 ± 1,11E+0	8,29E+0 ± 8,08E-1	7,71E+0 ± 2,02E-1	5,00E+0 ± 1,01E+0	1,10E+1 ± 1,41E+0	4,86E+0 ± 6,06E-1	6,64E+0 ± 3,03E-1	7,57E+0 ± 6,06E-1	6,29E+0 ± 4,04E-1	3,79E+0 ± 1,72E+0	3,07E+0 ± 1,52E+0
Cpu (15%)	5,86E+0 ± 4,04E-1	6,64E+0 ± 1,01E-1	1,02E+1 ± 3,03E-1	6,29E+0 ± 8,08E-1	6,71E+0 ± 6,06E-1	1,16E+1 ± 0,00E+0	5,71E+0 ± 1,41E+0	7,50E+0 ± 7,07E-1	5,50E+0 ± 5,05E-1	4,43E+0 ± 8,08E-1	2,57E+0 ± 6,06E-1	4,43E+0 ± 4,04E-1
Cpu (20%)	5,43E+0 ± 2,22E+0	5,57E+0 ± 4,04E-1	9,50E+0 ± 1,72E+0	6,29E+0 ± 1,62E+0	7,86E+0 ± 1,41E+0	1,18E+1 ± 3,03E-1	5,29E+0 ± 6,06E-1	7,79E+0 ± 1,11E+0	6,57E+0 ± 2,02E-1	5,14E+0 ± 0,00E+0	2,93E+0 ± 3,03E-1	3,57E+0 ± 8,08E-1
Cpu (25%)	5,64E+0 ± 9,09E-1	4,79E+0 ± 3,03E-1	9,57E+0 ± 0,00E+0	5,86E+0 ± 4,04E-1	8,64E+0 ± 7,07E-1	1,17E+1 ± 0,00E+0	6,86E+0 ± 2,02E-1	7,36E+0 ± 1,01E-1	5,50E+0 ± 3,03E-1	4,29E+0 ± 2,02E-1	2,07E+0 ± 1,01E-1	5,21E+0 ± 9,09E-1
Cpu (50%)	6,07E+0 ± 1,92E+0	6,21E+0 ± 5,05E-1	7,36E+0 ± 9,09E-1	7,14E+0 ± 1,21E+0	9,14E+0 ± 2,02E-1	1,09E+1 ± 2,02E-1	6,29E+0 ± 4,04E-1	8,86E+0 ± 4,04E-1	6,14E+0 ± 2,02E-1	4,43E+0 ± 2,02E-1	2,93E+0 ± 3,03E-1	2,07E+0 ± 9,09E-1
E. coli (1%)	6,36E+0 ± 1,92E+0	7,07E+0 ± 3,03E-1	5,00E+0 ± 0,00E+0	6,29E+0 ± 0,00E+0	4,21E+0 ± 9,09E-1	1,09E+1 ± 1,01E-1	9,79E+0 ± 5,05E-1	6,00E+0 ± 1,21E+0	6,71E+0 ± 1,01E+0	6,07E+0 ± 9,09E-1	3,86E+0 ± 2,02E-1	4,07E+0 ± 7,07E-1
E. coli (5%)	5,21E+0 ± 3,33E+0	5,71E+0 ± 2,02E-1	6,64E+0 ± 1,01E-1	7,64E+0 ± 1,31E+0	4,50E+0 ± 1,01E-1	1,10E+1 ± 2,02E-1	1,05E+1 ± 1,01E-1	8,36E+0 ± 7,07E-1	6,14E+0 ± 1,21E+0	5,07E+0 ± 9,09E-1	2,36E+0 ± 9,09E-1	3,93E+0 ± 2,12E+0
E. coli (10%)	8,14E+0 ± 2,02E-1	5,79E+0 ± 9,09E-1	6,07E+0 ± 3,03E-1	7,14E+0 ± 1,21E+0	3,36E+0 ± 9,09E-1	1,10E+1 ± 2,02E-1	1,03E+1 ± 2,02E-1	8,43E+0 ± 0,00E+0	6,29E+0 ± 0,00E+0	5,29E+0 ± 0,00E+0	2,14E+0 ± 0,00E+0	3,21E+0 ± 3,03E-1
E. coli (15%)	6,29E+0 ± 2,83E+0	5,00E+0 ± 8,08E-1	5,93E+0 ± 3,03E-1	6,64E+0 ± 7,07E-1	3,64E+0 ± 3,03E-1	1,04E+1 ± 9,09E-1	1,02E+1 ± 7,07E-1	8,57E+0 ± 2,02E-1	6,07E+0 ± 7,07E-1	4,86E+0 ± 8,08E-1	2,29E+0 ± 4,04E-1	3,14E+0 ± 6,06E-1
E. coli (20%)	7,36E+0 ± 2,73E+0	4,64E+0 ± 3,03E-1	6,07E+0 ± 1,31E+0	6,21E+0 ± 5,05E-1	4,29E+0 ± 1,21E+0	1,12E+1 ± 1,01E-1	1,07E+1 ± 2,02E-1	9,29E+0 ± 8,08E-1	6,71E+0 ± 6,06E-1	5,43E+0 ± 4,04E-1	2,50E+0 ± 3,03E-1	2,71E+0 ± 2,02E-1
E. coli (25%)	7,43E+0 ± 1,01E+0	4,64E+0 ± 1,01E-1	4,93E+0 ± 3,03E-1	5,79E+0 ± 9,09E-1	3,57E+0 ± 0,00E+0	9,71E+0 ± 0,00E+0	9,79E+0 ± 1,01E-1	7,79E+0 ± 3,03E-1	6,14E+0 ± 6,06E-1	4,86E+0 ± 1,01E+0	2,29E+0 ± 0,00E+0	1,43E+0 ± 0,00E+0
E. coli (50%)	7,79E+0 ± 1,01E-1	5,14E+0 ± 1,41E+0	5,50E+0 ± 5,05E-1	4,64E+0 ± 1,11E+0	7,00E+0 ± 8,08E-1	1,04E+1 ± 1,21E+0	1,00E+1 ± 2,02E-1	8,71E+0 ± 1,01E+0	5,43E+0 ± 6,06E-1	4,43E+0 ± 4,04E-1	1,93E+0 ± 1,01E-1	1,79E+0 ± 1,01E-1
Jap. v. (1%)	4,50E+0 ± 3,54E-1	4,42E+0 ± 0,00E+0	9,46E+0 ± 2,95E-1	8,58E+0 ± 0,00E+0	1,96E+0 ± 5,89E-2	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	8,67E+0 ± 0,00E+0	5,08E+0 ± 0,00E+0	4,92E+0 ± 0,00E+0	4,17E+0 ± 2,36E-1	1,67E+0 ± 4,71E-1
Jap. v. (5%)	6,21E+0 ± 2,95E-1	4,13E+0 ± 1,77E-1	9,33E+0 ± 1,18E-1	8,92E+0 ± 0,00E+0	3,08E+0 ± 0,00E+0	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	8,75E+0 ± 1,18E-1	6,42E+0 ± 1,18E-1	5,17E+0 ± 0,00E+0	1,83E+0 ± 1,18E-1	1,17E+0 ± 1,18E-1
Jap. v. (10%)	6,67E+0 ± 2,02E-1	4,17E+0 ± 9,09E-1	9,17E+0 ± 3,03E-1	9,08E+0 ± 1,21E+0	3,00E+0 ± 9,09E-1	1,10E+1 ± 2,02E-1	1,20E+1 ± 2,02E-1	8,75E+0 ± 0,00E+0	6,17E+0 ± 0,00E+0	5,00E+0 ± 0,00E+0	1,58E+0 ± 0,00E+0	1,42E+0 ± 3,03E-1
Jap. v. (15%)	6,88E+0 ± 1,77E-1	4,75E+0 ± 1,65E+0	7,75E+0 ± 1,89E+0	7,21E+0 ± 1,83E+0	4,54E+0 ± 1,24E+0	8,25E+0 ± 3,89E+0	9,04E+0 ± 4,18E+0	7,96E+0 ± 1,71E+0	6,63E+0 ± 7,66E-1	6,00E+0 ± 1,53E+0	4,50E+0 ± 4,24E+0	4,33E+0 ± 4,01E+0
Jap. v. (20%)	7,25E+0 ± 7,07E-1	4,75E+0 ± 1,89E+0	7,46E+0 ± 1,47E+0	7,17E+0 ± 1,06E+0	4,75E+0 ± 4,71E-1	8,13E+0 ± 4,07E+0	9,04E+0 ± 4,18E+0	8,17E+0 ± 1,89E+0	6,50E+0 ± 8,25E-1	5,88E+0 ± 1,83E+0	4,50E+0 ± 4,01E+0	4,29E+0 ± 4,18E+0
Jap. v. (25%)	6,83E+0 ± 1,18E-1	4,54E+0 ± 1,94E+0	7,21E+0 ± 1,00E+0	6,67E+0 ± 9,43E-1	5,88E+0 ± 8,84E-1	8,25E+0 ± 3,89E+0	8,96E+0 ± 4,30E+0	8,29E+0 ± 2,06E+0	6,54E+0 ± 1,00E+0	5,79E+0 ± 1,71E+0	4,54E+0 ± 4,42E+0	4,42E+0 ± 4,01E+0
Jap. v. (50%)	7,38E+0 ± 1,36E+0	4,71E+0 ± 1,59E+0	5,33E+0 ± 8,25E-1	4,75E+0 ± 2,36E-1	7,04E+0 ± 2,77E+0	8,33E+0 ± 3,77E+0	9,25E+0 ± 3,89E+0	7,46E+0 ± 3,01E+0	5,58E+0 ± 1,65E+0	4,96E+0 ± 6,48E-1	3,75E+0 ± 3,18E+0	3,71E+0 ± 3,12E+0
Lrs (1%)	5,80E+0 ± 6,36E-1	5,85E+0 ± 4,71E-2	5,56E+0 ± 1,89E-1	5,78E+0 ± 3,77E-1	6,75E+0 ± 3,77E-1	9,37E+0 ± 1,18E-1	9,60E+0 ± 7,07E-2	6,18E+0 ± 1,65E-1	6,03E+0 ± 2,83E-1	5,33E+0 ± 1,89E-1	4,32E+0 ± 1,18E-1	5,35E+0 ± 6,13E-1
Lrs (5%)	6,03E+0 ± 8,72E-1	5,54E+0 ± 8,72E-1	5,91E+0 ± 1,18E-1	5,67E+0 ± 4,71E-1	7,12E+0 ± 7,07E-1	1,06E+1 ± 1,41E-1	1,02E+1 ± 1,41E-1	5,79E+0 ± 6,13E-1	5,95E+0 ± 5,89E-1	4,60E+0 ± 6,84E-1	3,49E+0 ± 5,42E-1	5,72E+0 ± 8,25E-1
Lrs (10%)	5,58E+0 ± 3,60E-1	5,38E+0 ± 1,59E-1	6,01E+0 ± 5,13E-1	6,19E+0 ± 2,56E-1	7,56E+0 ± 1,32E-1	1,07E+1 ± 1,25E-1	9,93E+0 ± 2,70E-1	5,81E+0 ± 2,29E-1	6,08E+0 ± 1,94E-1	4,54E+0 ± 2,01E-1	3,61E+0 ± 1,73E-1	5,25E+0 ± 1,25E-1
Lrs (15%)	5,19E+0 ± 3,19E-1	5,26E+0 ± 2,29E-1	6,40E+0 ± 2,84E-1	6,63E+0 ± 3,74E-1	7,74E+0 ± 4,09E-1	1,10E+1 ± 1,59E-1	1,01E+1 ± 1,53E-1	5,48E+0 ± 8,32E-1	6,15E+0 ± 5,20E-1	4,55E+0 ± 3,81E-1	3,71E+0 ± 8,73E-1	4,36E+0 ± 8,04E-1
Lrs (20%)	4,82E+0 ± 4,16E-2	4,47E+0 ± 4,78E-1	6,86E+0 ± 5,34E-1	7,49E+0 ± 4,30E-1	7,81E+0 ± 1,32E-1	1,08E+1 ± 2,77E-2	9,86E+0 ± 1,25E-1	5,25E+0 ± 8,32E-2	6,38E+0 ± 2,56E-1	4,65E+0 ± 1,80E-1	4,11E+0 ± 3,33E-1	3,95E+0 ± 4,23E-1
Lrs (25%)	4,20E+0 ± 4,37E-1	4,32E+0 ± 4,16E-2	7,81E+0 ± 3,88E-1	8,26E+0 ± 1,25E-1	7,72E+0 ± 6,24E-2	1,08E+1 ± 9,01E-2	9,66E+0 ± 1,11E-1	5,34E+0 ± 3,05E-1	6,23E+0 ± 1,39E-2	4,34E+0 ± 4,16E-2	3,83E+0 ± 1,11E-1	3,91E+0 ± 6,24E-2
Lrs (50%)	2,82E+0 ± 9,01E-2	4,75E+0 ± 2,22E-1	1,02E+1 ± 3,47E-2	1,09E+1 ± 1,39E-2	5,90E+0 ± 1,18E-1	8,08E+0 ± 2,98E-1	7,46E+0 ± 2,08E-2	4,54E+0 ± 1,46E-1	8,03E+0 ± 3,47E-2	5,98E+0 ± 9,01E-2	5,85E+0 ± 9,71E-2	1,92E+0 ± 1,39E-1
M. mas. (1%)	6,42E+0 ± 2,00E+0	5,92E+0 ± 1,18E-1	5,00E+0 ± 9,43E-1	7,58E+0 ± 5,89E-1	3,75E+0 ± 1,06E+0	1,13E+1 ± 0,00E+0	8,58E+0 ± 1,18E-1	6,58E+0 ± 1,30E+0	7,25E+0 ± 5,89E-1	6,17E+0 ± 2,36E-1	3,58E+0 ± 8,25E-1	4,92E+0 ± 1,30E+0
M. mas. (5%)	7,25E+0 ± 1,06E+0	4,58E+0 ± 1,53E+0	5,25E+0 ± 1,77E+0	7,50E+0 ± 9,43E-1	3,83E+0 ± 4,71E-1	1,15E+1 ± 0,00E+0	9,17E+0 ± 4,71E-1	8,75E+0 ± 8,25E-1	8,00E+0 ± 9,43E-1	6,50E+0 ± 9,43E-1	3,92E+0 ± 1,18E-1	1,17E+0 ± 2,36E-1
M. mas. (10%)	6,83E+0 ± 2,36E+0	6,33E+0 ± 4,71E-1	4,00E+0 ± 7,07E-1	7,83E+0 ± 2,36E-1	4,25E+0 ± 1,30E+0	1,15E+1 ± 0,00E+0	8,67E+0 ± 0,00E+0	8,83E+0 ± 1,65E+0	7,50E+0 ± 2,36E-1	6,58E+0 ± 3,54E-1	3,83E+0 ± 0,00E+0	1,25E+0 ± 1,18E-1

Po. m.	RD	MD	LMS	LR	KNN	Pov.	ZeroR	Bag.	RRG	RRG1	RRG2	SPSS
M. mas. (15%)	6,67E+0 ± 1,89E+0	5,75E+0 ± 1,18E-1	4,33E+0 ± 2,36E-1	8,00E+0 ± 1,18E+0	3,75E+0 ± 3,54E-1	1,15E+1 ± 0,00E+0	8,42E+0 ± 1,18E-1	8,92E+0 ± 5,89E-1	8,42E+0 ± 1,18E-1	7,00E+0 ± 4,71E-1	3,92E+0 ± 3,54E-1	1,08E+0 ± 1,18E-1
M. mas. (20%)	7,67E+0 ± 2,36E-1	4,58E+0 ± 5,89E-1	5,67E+0 ± 0,00E+0	6,33E+0 ± 4,71E-1	5,42E+0 ± 8,25E-1	1,15E+1 ± 0,00E+0	8,00E+0 ± 4,71E-1	1,01E+1 ± 1,18E-1	7,33E+0 ± 4,71E-1	6,33E+0 ± 4,71E-1	3,17E+0 ± 0,00E+0	1,42E+0 ± 1,18E-1
M. mas. (25%)	8,58E+0 ± 3,54E-1	4,83E+0 ± 1,41E+0	5,25E+0 ± 5,89E-1	5,08E+0 ± 1,30E+0	4,50E+0 ± 2,36E-1	1,15E+1 ± 0,00E+0	8,17E+0 ± 2,36E-1	1,02E+1 ± 2,36E-1	7,67E+0 ± 2,36E-1	6,58E+0 ± 3,54E-1	3,42E+0 ± 1,18E-1	1,67E+0 ± 2,36E-1
M. mas. (50%)	9,33E+0 ± 2,36E-1	3,75E+0 ± 5,89E-1	7,00E+0 ± 1,41E+0	3,58E+0 ± 1,53E+0	6,08E+0 ± 5,89E-1	1,13E+1 ± 2,36E-1	8,67E+0 ± 1,41E+0	9,50E+0 ± 0,00E+0	6,42E+0 ± 5,89E-1	5,50E+0 ± 7,07E-1	1,92E+0 ± 3,54E-1	4,50E+0 ± 1,41E+0
Ozone (1%)	4,80E+0 ± 1,87E-1	4,82E+0 ± 1,57E-1	5,58E+0 ± 1,28E-1	4,71E+0 ± 3,93E-2	6,79E+0 ± 4,12E-1	1,03E+1 ± 1,96E-2	1,15E+1 ± 8,84E-2	5,34E+0 ± 3,44E-1	5,05E+0 ± 2,95E-2	4,41E+0 ± 2,95E-2	3,31E+0 ± 5,89E-2	1,00E+1 ± 2,26E-1
Ozone (5%)	5,41E+0 ± 1,87E-1	4,33E+0 ± 7,86E-2	5,30E+0 ± 4,91E-2	5,07E+0 ± 2,55E-1	8,77E+0 ± 9,82E-3	1,10E+1 ± 2,95E-2	1,16E+1 ± 9,82E-3	6,06E+0 ± 2,46E-1	4,86E+0 ± 0,00E+0	3,56E+0 ± 1,28E-1	2,44E+0 ± 1,28E-1	8,74E+0 ± 2,95E-2
Ozone (10%)	4,97E+0 ± 6,87E-2	3,97E+0 ± 5,89E-2	5,94E+0 ± 7,86E-2	6,22E+0 ± 2,06E-1	8,94E+0 ± 4,91E-2	1,12E+1 ± 9,82E-3	1,16E+1 ± 9,82E-3	5,73E+0 ± 1,28E-1	5,31E+0 ± 1,47E-1	3,69E+0 ± 8,84E-2	2,99E+0 ± 2,95E-2	6,67E+0 ± 1,08E-1
Ozone (15%)	5,03E+0 ± 2,65E-1	3,26E+0 ± 3,93E-2	6,28E+0 ± 9,82E-2	6,57E+0 ± 7,86E-2	9,23E+0 ± 3,63E-1	1,12E+1 ± 0,00E+0	1,16E+1 ± 1,96E-2	6,58E+0 ± 1,87E-1	5,55E+0 ± 9,82E-3	3,88E+0 ± 0,00E+0	2,49E+0 ± 3,93E-2	5,70E+0 ± 1,47E-1
Ozone (20%)	5,07E+0 ± 1,96E-1	3,05E+0 ± 2,26E-1	6,94E+0 ± 3,04E-1	7,63E+0 ± 3,04E-1	9,38E+0 ± 1,67E-1	1,12E+1 ± 0,00E+0	1,14E+1 ± 4,91E-2	6,49E+0 ± 1,18E-1	5,40E+0 ± 0,00E+0	3,56E+0 ± 7,86E-2	2,67E+0 ± 2,95E-2	4,36E+0 ± 9,82E-2
Ozone (25%)	4,57E+0 ± 9,82E-2	3,07E+0 ± 2,55E-1	7,79E+0 ± 7,86E-2	8,48E+0 ± 6,87E-2	9,30E+0 ± 4,91E-2	1,12E+1 ± 0,00E+0	1,12E+1 ± 9,82E-2	5,94E+0 ± 1,37E-1	5,38E+0 ± 1,57E-1	3,45E+0 ± 2,95E-2	2,92E+0 ± 9,82E-3	3,76E+0 ± 5,89E-2
Ozone (50%)	4,06E+0 ± 2,95E-2	3,49E+0 ± 3,63E-1	1,02E+1 ± 2,95E-2	1,09E+1 ± 4,91E-2	7,09E+0 ± 8,84E-2	9,31E+0 ± 1,28E-1	9,31E+0 ± 1,08E-1	5,66E+0 ± 9,82E-3	6,69E+0 ± 3,93E-2	4,68E+0 ± 0,00E+0	4,49E+0 ± 1,96E-2	1,23E+0 ± 4,91E-2
P. bl. (1%)	4,91E+0 ± 1,03E+0	6,23E+0 ± 1,48E+0	9,50E+0 ± 3,21E-1	6,59E+0 ± 9,64E-1	5,27E+0 ± 1,29E-1	1,00E+1 ± 2,19E+0	8,00E+0 ± 1,67E+0	5,64E+0 ± 9,00E-1	5,36E+0 ± 5,14E-1	5,14E+0 ± 4,50E-1	4,68E+0 ± 5,79E-1	4,64E+0 ± 2,19E+0
P. bl. (5%)	5,27E+0 ± 9,00E-1	4,91E+0 ± 1,29E-1	8,41E+0 ± 1,09E+0	7,09E+0 ± 3,86E-1	6,23E+0 ± 1,09E+0	1,10E+1 ± 4,50E-1	9,73E+0 ± 1,29E-1	5,95E+0 ± 5,79E-1	5,91E+0 ± 3,86E-1	5,32E+0 ± 3,21E-1	3,73E+0 ± 2,57E-1	3,09E+0 ± 3,86E-1
P. bl. (10%)	5,14E+0 ± 3,21E-1	5,59E+0 ± 4,50E-1	9,50E+0 ± 1,74E+0	6,45E+0 ± 6,43E-1	6,64E+0 ± 5,14E-1	1,10E+1 ± 4,50E-1	1,00E+1 ± 5,79E-1	6,36E+0 ± 2,57E-1	4,95E+0 ± 5,79E-1	4,23E+0 ± 4,50E-1	2,86E+0 ± 3,21E-1	3,68E+0 ± 3,21E-1
P. bl. (15%)	4,82E+0 ± 2,57E-1	4,14E+0 ± 7,07E-1	8,64E+0 ± 1,41E+0	6,91E+0 ± 1,54E+0	7,59E+0 ± 1,22E+0	1,14E+1 ± 1,93E-1	1,04E+1 ± 5,14E-1	5,82E+0 ± 3,86E-1	5,68E+0 ± 4,50E-1	4,86E+0 ± 6,43E-2	3,18E+0 ± 2,57E-1	3,32E+0 ± 1,93E-1
P. bl. (20%)	5,50E+0 ± 1,86E+0	4,95E+0 ± 1,09E+0	9,59E+0 ± 4,50E-1	5,91E+0 ± 9,00E-1	7,00E+0 ± 3,86E-1	1,12E+1 ± 3,21E-1	1,02E+1 ± 3,21E-1	7,41E+0 ± 5,79E-1	4,82E+0 ± 6,43E-1	3,73E+0 ± 7,71E-1	2,59E+0 ± 7,07E-1	3,77E+0 ± 6,43E-2
P. bl. (25%)	5,86E+0 ± 4,50E-1	4,73E+0 ± 2,57E-1	8,95E+0 ± 3,21E-1	5,68E+0 ± 6,43E-2	8,64E+0 ± 1,29E-1	1,13E+1 ± 2,57E-1	1,02E+1 ± 1,93E-1	7,00E+0 ± 3,86E-1	5,09E+0 ± 2,57E-1	4,14E+0 ± 4,50E-1	2,41E+0 ± 3,21E-1	3,14E+0 ± 3,21E-1
P. bl. (50%)	5,27E+0 ± 1,03E+0	6,09E+0 ± 5,14E-1	7,36E+0 ± 1,29E-1	8,09E+0 ± 3,86E-1	8,64E+0 ± 2,57E-1	1,03E+1 ± 3,21E-1	8,14E+0 ± 8,36E-1	6,27E+0 ± 6,43E-1	6,82E+0 ± 3,86E-1	4,73E+0 ± 6,43E-1	3,77E+0 ± 9,64E-1	1,77E+0 ± 1,93E-1
P. Ind. (1%)	6,44E+0 ± 6,29E-1	6,06E+0 ± 1,02E+0	7,39E+0 ± 1,18E+0	7,28E+0 ± 1,18E+0	4,67E+0 ± 9,43E-1	1,16E+1 ± 7,86E-2	6,06E+0 ± 1,49E+0	6,83E+0 ± 2,36E-1	7,17E+0 ± 8,64E-1	6,28E+0 ± 5,50E-1	3,11E+0 ± 3,14E-1	4,22E+0 ± 3,14E-1
P. Ind. (5%)	6,61E+0 ± 2,36E-1	5,22E+0 ± 0,00E+0	7,44E+0 ± 0,00E+0	8,28E+0 ± 2,36E-1	4,78E+0 ± 1,57E-1	1,17E+1 ± 0,00E+0	5,78E+0 ± 0,00E+0	8,17E+0 ± 7,86E-2	8,17E+0 ± 2,36E-1	6,94E+0 ± 2,36E-1	3,28E+0 ± 5,50E-1	1,56E+0 ± 6,29E-1
P. Ind. (10%)	8,28E+0 ± 3,93E-1	5,67E+0 ± 6,29E-1	6,44E+0 ± 6,29E-1	7,78E+0 ± 1,57E-1	4,83E+0 ± 1,34E+0	1,17E+1 ± 0,00E+0	5,50E+0 ± 7,86E-2	9,39E+0 ± 3,93E-1	7,61E+0 ± 7,86E-2	6,22E+0 ± 1,57E-1	2,94E+0 ± 5,50E-1	1,33E+0 ± 0,00E+0
P. Ind. (15%)	4,33E+0 ± 9,43E-1	4,06E+0 ± 2,36E-1	7,39E+0 ± 3,93E-1	8,22E+0 ± 1,57E-1	6,39E+0 ± 7,86E-2	1,17E+1 ± 0,00E+0	5,83E+0 ± 7,86E-2	1,02E+1 ± 3,14E-1	8,00E+0 ± 3,14E-1	6,56E+0 ± 4,71E-1	3,89E+0 ± 3,14E-1	1,28E+0 ± 2,36E-1
P. Ind. (20%)	5,94E+0 ± 5,50E-1	4,33E+0 ± 1,57E-1	7,39E+0 ± 5,50E-1	7,67E+0 ± 3,14E-1	6,94E+0 ± 2,91E+0	1,16E+1 ± 0,00E+0	5,67E+0 ± 4,71E-1	9,89E+0 ± 4,71E-1	7,56E+0 ± 7,86E-1	6,11E+0 ± 4,71E-1	3,50E+0 ± 3,93E-1	1,11E+0 ± 0,00E+0
P. Ind. (25%)	7,61E+0 ± 2,75E+0	4,39E+0 ± 5,50E-1	6,94E+0 ± 7,07E-1	6,83E+0 ± 8,64E-1	6,56E+0 ± 1,10E+0	1,17E+1 ± 0,00E+0	5,67E+0 ± 0,00E+0	1,02E+1 ± 3,93E-1	7,50E+0 ± 7,07E-1	6,11E+0 ± 9,43E-1	3,06E+0 ± 2,36E-1	1,17E+0 ± 7,86E-2
P. Ind. (50%)	9,94E+0 ± 7,86E-2	4,39E+0 ± 2,36E-1	7,28E+0 ± 7,86E-2	4,61E+0 ± 7,86E-2	7,94E+0 ± 5,50E-1	1,17E+1 ± 0,00E+0	7,17E+0 ± 7,86E-2	1,03E+1 ± 7,86E-2	6,17E+0 ± 5,50E-1	4,83E+0 ± 7,86E-2	2,33E+0 ± 0,00E+0	1,00E+0 ± 0,00E+0
Segm. (1%)	6,65E+0 ± 1,67E+0	6,40E+0 ± 6,47E-2	5,62E+0 ± 2,17E-1	5,96E+0 ± 6,45E-1	5,58E+0 ± 8,25E-1	1,06E+1 ± 4,99E-1	5,42E+0 ± 5,89E-1	6,56E+0 ± 4,62E-1	6,13E+0 ± 8,13E-1	5,36E+0 ± 7,42E-1	3,41E+0 ± 6,63E-1	5,42E+0 ± 1,07E+0
Segm. (5%)	6,47E+0 ± 7,44E-2	5,16E+0 ± 2,23E-1	7,13E+0 ± 4,84E-1	6,32E+0 ± 1,49E-1	5,61E+0 ± 3,35E-1	1,08E+1 ± 1,49E-1	6,42E+0 ± 4,47E-1	6,61E+0 ± 1,12E-1	6,53E+0 ± 4,47E-1	5,16E+0 ± 5,21E-1	2,82E+0 ± 4,84E-1	5,18E+0 ± 8,56E-1
Segm. (10%)	5,79E+0 ± 7,44E-2	6,13E+0 ± 4,84E-1	6,66E+0 ± 4,84E-1	6,63E+0 ± 3,72E-1	5,71E+0 ± 1,86E-1	1,03E+1 ± 1,86E-1	6,66E+0 ± 7,07E-1	5,95E+0 ± 8,93E-1	6,50E+0 ± 4,84E-1	4,87E+0 ± 5,58E-1	3,26E+0 ± 1,49E-1	5,45E+0 ± 1,15E+0
Segm. (15%)	4,92E+0 ± 6,33E-1	5,71E+0 ± 9,30E-1	6,92E+0 ± 4,09E-1	6,13E+0 ± 2,61E-1	7,03E+0 ± 9,30E-1	1,09E+1 ± 1,12E-1	7,50E+0 ± 3,72E-2	6,97E+0 ± 1,86E-1	6,42E+0 ± 6,70E-1	5,00E+0 ± 5,95E-1	3,18E+0 ± 4,09E-1	3,21E+0 ± 2,98E-1
Segm. (20%)	4,89E+0 ± 1,86E+0	5,76E+0 ± 9,30E-1	6,61E+0 ± 3,72E-2	6,32E+0 ± 2,23E-1	7,16E+0 ± 4,47E-1	1,09E+1 ± 7,44E-2	8,00E+0 ± 1,04E+0	6,26E+0 ± 1,49E-1	6,26E+0 ± 2,98E-1	4,82E+0 ± 3,35E-1	3,13E+0 ± 1,12E-1	3,76E+0 ± 1,12E-1
Segm. (25%)	6,63E+0 ± 7,44E-1	5,47E+0 ± 5,95E-1	6,24E+0 ± 3,72E-2	6,26E+0 ± 5,95E-1	6,92E+0 ± 4,84E-1	1,10E+1 ± 3,72E-2	7,92E+0 ± 2,61E-1	7,26E+0 ± 2,23E-1	5,87E+0 ± 2,61E-1	4,50E+0 ± 1,12E-1	2,84E+0 ± 1,49E-1	3,08E+0 ± 1,86E-1
Segm. (50%)	4,53E+0 ± 1,27E+0	7,26E+0 ± 2,23E-1	7,87E+0 ± 9,30E-1	8,13E+0 ± 1,30E+0	6,03E+0 ± 1,86E-1	8,42E+0 ± 0,00E+0	5,66E+0 ± 3,72E-2	5,92E+0 ± 1,12E-1	7,16E+0 ± 2,23E-1	5,39E+0 ± 2,61E-1	4,66E+0 ± 3,72E-2	2,68E+0 ± 7,44E-1

Po. m.	RD	MD	LMS	LR	KNN	Pov.	ZeroR	Bag.	RRG	RRG1	RRG2	SPSS
Spell. (1%)	6,96E+0 ± 8,42E-2	3,17E+0 ± 2,32E+0	5,65E+0 ± 3,20E-1	3,22E+0 ± 1,06E+0	8,58E+0 ± 2,09E+0	1,03E+1 ± 7,41E-1	1,20E+1 ± 2,22E+0	8,09E+0 ± 1,03E+0	2,88E+0 ± 1,11E+0	2,88E+0 ± 6,90E-1	2,43E+0 ± 3,87E-1	9,58E+0 ± 3,03E-1
Spell. (5%)	7,88E+0 ± 2,02E-1	3,49E+0 ± 1,62E+0	7,03E+0 ± 8,08E-1	4,53E+0 ± 6,06E-1	9,58E+0 ± 2,02E-1	1,10E+1 ± 6,06E-1	1,20E+1 ± 3,03E-1	8,95E+0 ± 1,11E+0	4,23E+0 ± 7,07E-1	3,49E+0 ± 3,03E-1	2,36E+0 ± 4,04E-1	2,61E+0 ± 0,00E+0
Spell. (10%)	8,05E+0 ± 1,72E+0	4,14E+0 ± 1,11E+0	7,04E+0 ± 8,08E-1	4,65E+0 ± 2,02E-1	9,77E+0 ± 1,01E+0	1,10E+1 ± 1,41E+0	1,20E+1 ± 6,06E-1	9,03E+0 ± 3,03E-1	4,92E+0 ± 6,06E-1	3,61E+0 ± 4,04E-1	2,01E+0 ± 1,72E+0	1,49E+0 ± 1,52E+0
Spell. (15%)	8,19E+0 ± 3,67E-2	4,34E+0 ± 8,26E-2	6,84E+0 ± 9,18E-2	4,30E+0 ± 2,20E-1	9,71E+0 ± 3,67E-2	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	8,99E+0 ± 3,67E-2	5,06E+0 ± 1,19E-1	3,68E+0 ± 9,18E-3	1,88E+0 ± 5,51E-2	1,79E+0 ± 2,11E-1
Spell. (20%)	8,26E+0 ± 7,35E-2	4,19E+0 ± 8,26E-2	6,60E+0 ± 1,19E-1	4,19E+0 ± 5,51E-2	9,66E+0 ± 2,75E-2	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,08E+0 ± 3,67E-2	5,18E+0 ± 9,18E-2	3,68E+0 ± 1,74E-1	1,92E+0 ± 8,26E-2	1,90E+0 ± 2,30E-1
Spell. (25%)	8,27E+0 ± 3,67E-2	4,36E+0 ± 1,74E-1	6,21E+0 ± 1,74E-1	4,14E+0 ± 8,26E-2	9,54E+0 ± 1,01E-1	1,10E+1 ± 0,00E+0	1,20E+1 ± 0,00E+0	9,18E+0 ± 6,43E-2	5,17E+0 ± 9,18E-2	3,71E+0 ± 4,59E-2	2,04E+0 ± 5,51E-2	2,03E+0 ± 9,18E-3
Spell. (50%)	7,75E+0 ± 6,43E-2	4,42E+0 ± 1,99E+0	4,66E+0 ± 5,88E-1	6,28E+0 ± 8,26E-2	8,31E+0 ± 6,61E-1	1,05E+1 ± 2,94E-1	1,20E+1 ± 0,00E+0	8,79E+0 ± 7,25E-1	5,86E+0 ± 1,01E-1	4,14E+0 ± 1,65E-1	2,86E+0 ± 4,59E-1	1,79E+0 ± 1,65E-1
Wisc. (1%)	5,82E+0 ± 6,36E-1	5,63E+0 ± 4,71E-2	6,20E+0 ± 1,89E-1	5,50E+0 ± 3,77E-1	6,77E+0 ± 3,77E-1	1,09E+1 ± 1,18E-1	9,12E+0 ± 7,07E-2	6,25E+0 ± 1,65E-1	5,77E+0 ± 2,83E-1	5,17E+0 ± 1,89E-1	3,88E+0 ± 1,18E-1	5,77E+0 ± 6,13E-1
Wisc. (5%)	6,12E+0 ± 8,72E-1	5,92E+0 ± 8,72E-1	6,45E+0 ± 1,18E-1	5,00E+0 ± 4,71E-1	8,17E+0 ± 7,07E-1	1,13E+1 ± 1,41E-1	9,57E+0 ± 1,41E-1	6,93E+0 ± 6,13E-1	6,08E+0 ± 5,89E-1	4,65E+0 ± 6,84E-1	2,92E+0 ± 5,42E-1	4,38E+0 ± 8,25E-1
Wisc. (10%)	6,63E+0 ± 9,43E-2	4,95E+0 ± 7,07E-2	5,85E+0 ± 4,48E-1	4,75E+0 ± 1,15E+0	9,05E+0 ± 6,84E-1	1,12E+1 ± 0,00E+0	9,67E+0 ± 1,89E-1	7,03E+0 ± 8,96E-1	5,72E+0 ± 2,59E-1	4,30E+0 ± 1,41E-1	2,88E+0 ± 2,59E-1	5,32E+0 ± 1,01E+0
Wisc. (15%)	5,87E+0 ± 5,66E-1	4,38E+0 ± 2,59E-1	6,08E+0 ± 1,18E-1	5,60E+0 ± 2,36E-1	9,58E+0 ± 2,36E-2	1,14E+1 ± 0,00E+0	9,62E+0 ± 2,12E-1	7,40E+0 ± 2,36E-1	6,08E+0 ± 7,07E-2	4,63E+0 ± 4,71E-2	2,48E+0 ± 7,07E-2	4,28E+0 ± 9,19E-1
Wisc. (20%)	5,98E+0 ± 1,18E-1	4,83E+0 ± 2,83E-1	5,97E+0 ± 9,43E-2	6,33E+0 ± 9,43E-1	9,10E+0 ± 2,83E-1	1,14E+1 ± 2,36E-2	9,58E+0 ± 7,07E-2	7,65E+0 ± 2,12E-1	5,77E+0 ± 0,00E+0	4,17E+0 ± 9,43E-2	2,57E+0 ± 9,43E-2	3,97E+0 ± 1,41E-1
Wisc. (25%)	5,93E+0 ± 4,71E-2	4,43E+0 ± 1,41E-1	6,43E+0 ± 5,66E-1	6,17E+0 ± 4,24E-1	9,52E+0 ± 4,01E-1	1,15E+1 ± 0,00E+0	8,97E+0 ± 1,41E-1	7,53E+0 ± 3,77E-1	6,10E+0 ± 9,43E-2	4,47E+0 ± 0,00E+0	2,65E+0 ± 3,06E-1	3,75E+0 ± 2,59E-1
Wisc. (50%)	5,60E+0 ± 1,41E-1	4,70E+0 ± 6,13E-1	8,27E+0 ± 2,36E-1	9,58E+0 ± 7,07E-2	8,08E+0 ± 1,65E-1	1,02E+1 ± 2,59E-1	5,95E+0 ± 1,18E-1	7,13E+0 ± 9,43E-2	6,53E+0 ± 2,36E-1	4,63E+0 ± 2,83E-1	4,18E+0 ± 1,65E-1	2,15E+0 ± 1,18E-1
Yeast (1%)	5,94E+0 ± 3,45E+0	5,81E+0 ± 1,33E+0	3,94E+0 ± 7,95E-1	7,19E+0 ± 9,72E-1	2,63E+0 ± 3,54E-1	1,10E+1 ± 0,00E+0	9,69E+0 ± 8,84E-2	7,63E+0 ± 1,06E+0	7,81E+0 ± 8,84E-2	7,25E+0 ± 0,00E+0	5,06E+0 ± 4,42E-1	1,63E+0 ± 8,84E-1
Yeast (5%)	7,50E+0 ± 0,00E+0	6,50E+0 ± 3,54E-1	5,81E+0 ± 8,84E-2	7,94E+0 ± 8,84E-2	2,44E+0 ± 2,65E-1	1,09E+1 ± 8,84E-2	1,04E+1 ± 3,54E-1	6,81E+0 ± 2,65E-1	8,19E+0 ± 8,84E-2	6,56E+0 ± 7,95E-1	3,19E+0 ± 2,65E-1	1,00E+0 ± 0,00E+0
Yeast (10%)	5,63E+0 ± 3,36E+0	5,31E+0 ± 2,65E-1	6,31E+0 ± 7,95E-1	8,31E+0 ± 1,15E+0	3,44E+0 ± 1,15E+0	1,11E+1 ± 1,77E-1	1,08E+1 ± 3,54E-1	8,56E+0 ± 6,19E-1	7,56E+0 ± 7,95E-1	6,13E+0 ± 7,07E-1	3,44E+0 ± 7,95E-1	1,00E+0 ± 0,00E+0
Yeast (15%)	8,19E+0 ± 6,19E-1	6,81E+0 ± 2,21E+0	4,88E+0 ± 2,30E+0	7,63E+0 ± 1,77E-1	2,63E+0 ± 1,77E-1	1,09E+1 ± 0,00E+0	1,04E+1 ± 3,54E-1	8,06E+0 ± 2,65E-1	7,56E+0 ± 6,19E-1	6,06E+0 ± 2,65E-1	3,13E+0 ± 1,77E-1	1,00E+0 ± 0,00E+0
Yeast (20%)	8,50E+0 ± 5,30E-1	5,50E+0 ± 1,77E-1	4,50E+0 ± 7,07E-1	8,06E+0 ± 6,19E-1	2,88E+0 ± 5,30E-1	1,12E+1 ± 8,84E-2	1,07E+1 ± 8,84E-2	8,81E+0 ± 7,95E-1	7,56E+0 ± 4,42E-1	5,69E+0 ± 4,42E-1	3,25E+0 ± 1,77E-1	1,00E+0 ± 0,00E+0
Yeast (25%)	9,00E+0 ± 8,84E-1	6,25E+0 ± 3,54E-1	5,63E+0 ± 5,30E-1	7,19E+0 ± 7,95E-1	3,81E+0 ± 2,65E-1	1,11E+1 ± 8,84E-2	1,07E+1 ± 8,84E-2	9,44E+0 ± 8,84E-2	6,06E+0 ± 4,42E-1	4,31E+0 ± 4,42E-1	2,88E+0 ± 1,77E-1	1,13E+0 ± 0,00E+0
Yeast (50%)	9,25E+0 ± 3,54E-1	6,31E+0 ± 2,65E-1	5,25E+0 ± 2,83E+0	6,44E+0 ± 6,19E-1	5,19E+0 ± 9,72E-1	1,08E+1 ± 1,77E-1	9,50E+0 ± 0,00E+0	8,56E+0 ± 8,84E-2	6,31E+0 ± 2,65E-1	4,94E+0 ± 2,65E-1	2,63E+0 ± 3,54E-1	1,75E+0 ± 7,07E-1
Povp. rang*	**6,61**	**4,36**	**7,77**	**7,37**	**6,61**	**11,65**	**9,96**	**8,83**	**6,74**	**4,12**	**1,72**	**2,27**

*) Povprečni rang po Friedmanovem testu (N=98, χ^2=734,202, sp=11, sig.=0,000)

Kakor pri razvrščanju metod glede natančnosti smo tudi pri primerjanju uspešnosti ohranjanja variance uporabili Friedmanov test (Tabela 6.1-20), ki je potrdil našo domnevo, da se metode po sposobnosti ohranjanja variance signifikantno razlikujejo med seboj. Najbolje rangirana metoda je različica našega rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance, ki ji sledi metoda orodja za statistično analizo SPSS in nato še različica rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance. Osnovni rotacijski regresijski gozd (brez dodatne metode za ohranjanje variance) se je v tej razvrstitvi znašel prav na sredi.

Tabela 6.1-20: Friedmanov test za primerjavo metod po sposobnosti ohranjanja variance

N	χ^2	stopnje prostosti	p
98	734,202	11	0,000

Z računanjem povprečnih rangov metod že v fazi ocenjevanja ohranjanja variance pri posameznih poskusih smo žrtvovali natančnost splošne ocene. Upoštevati je namreč treba, da večina metod dobro ohranja varianco pri manjših odstotkih manjkajočih vrednosti, kar je posledica omejenega vpliva napovedanih vrednosti na skupno varianco vrednosti atributa, če je njihovo število majhno v primerjavi z vsemi vrednostmi atributa. Zaradi tega je smiselno primerjati metode tudi samo pri večjih deležih manjkajočih vrednosti, kjer lahko pričakujemo, da so razlike med posameznimi metodami bolj izrazite. S tem namenom smo iz seznama meritev odstranili vse poskuse, kjer delež manjkajočih vrednosti ni presegel 20%. Na preostalih 28 primerih smo znova uporabili Friedmanov test, ki je spet potrdil signifikantno razlikovanje metod (Tabela 6.1-21), pri čemer pa je bila najbolje rangirana metoda orodja za statistično analizo SPSS (Tabela 6.1-22). Naš rotacijski regresijski gozd z agresivno metodo ohranjanja variance se je uvrstil na drugo mesto, sledila mu je različica z neagresivno metodo, medtem ko se je različica brez dodatne metode za ohranjanje variance izkazala podobno, oziroma le malo bolje, kot večina preostalih metod (Graf 6.1-19).

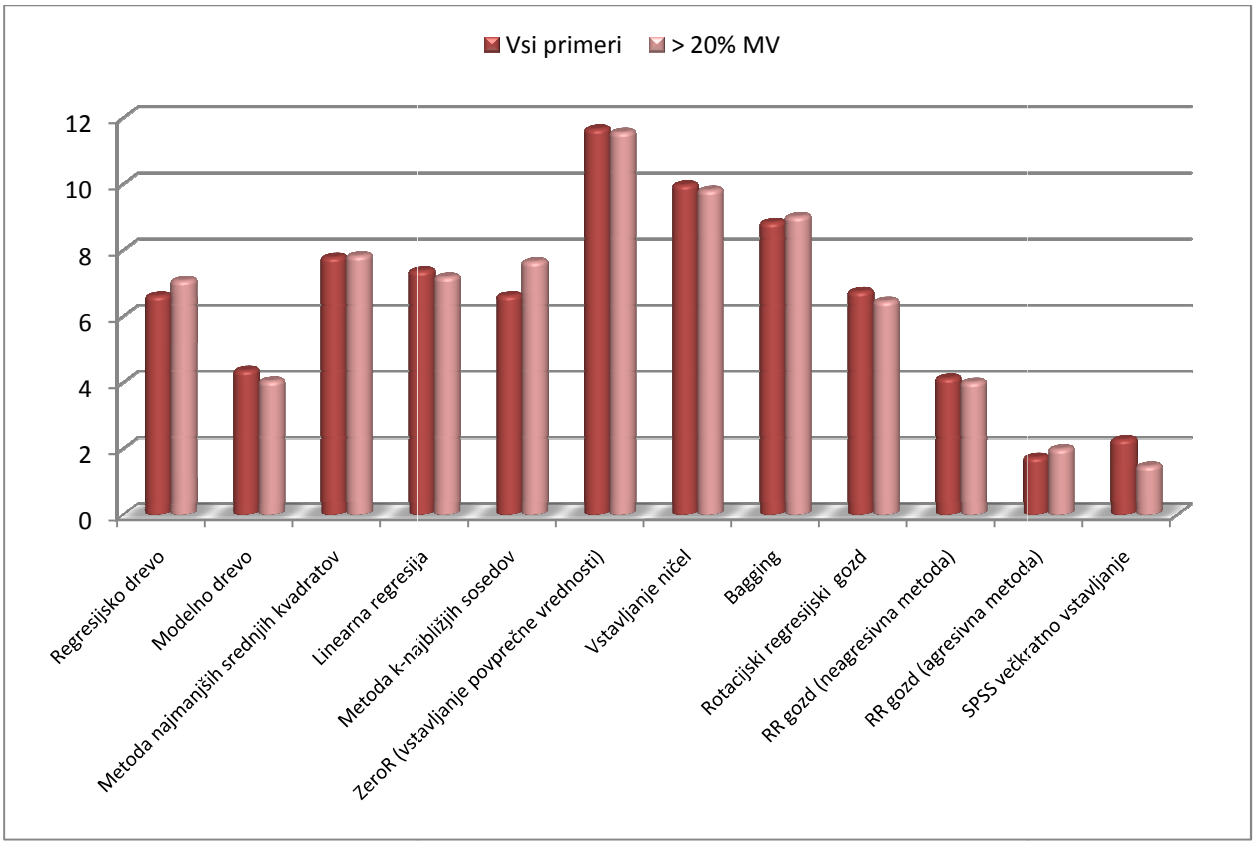
Tabela 6.1-21: Friedmanov test za primerjavo metod po sposobnosti ohranjanja variance pri 25% in 50% manjkajočih vrednosti

N	χ^2	stopnje prostosti	p
28	225,455	11	0,000

Tabela 6.1-22: Friedmanova razvrstitev metod po rangih glede na uspešnost ohranjanja variance na podatkovnih množicah z več kot 20% manjkajočih vrednosti

	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Povprečni rang*	7,07	4,04	7,82	7,18	7,64	11,55	9,80	9,00	6,45	4,00	1,98	1,46

*) Povprečni rang po Friedmanovem testu (N=28, χ^2=225,455, sp=11, sig.=0,000)



Graf 6.1-19: Povprečni rangi metod za napovedovanje manjkajočih vrednosti glede na njihovo sposobnost ohranjanja variance pri vseh deležih manjkajočih vrednosti in pri > 20% manjkajočih vrednosti

Za preverjanje ničelne hipoteze, da se dve primerjani metodi bistveno ne razlikujeta po svoji sposobnosti ohranjanja variance, smo (tako kot pri primerjanju metod po natančnosti) uporabili Wilcoxonov test predznačenih rangov. Na podlagi povprečnih rangov smo izbrali 3 metode, ki so po svoji sposobnosti ohranjanja variance najbliže rotacijskemu regresijskemu gozdu z agresivno metodo za ohranjanje variance: metoda orodja za statistično analizo SPSS, rotacijski regresijski gozd z neagresivno metodo za ohranjanje variance in modelno drevo. Vsako izmed teh metod smo s pomočjo Wilcoxonovega testa primerjali z rotacijskim regresijskim gozdom z agresivno metodo za ohranjanje variance. Primerjave smo opravili dvakrat, prvič na vseh primerih podatkovnih množic in drugič na primerih, kjer je bil delež manjkajočih vrednosti večji kot 20%. Rezultati analize, opravljene z orodjem za statistično analizo SPSS, so predstavljeni v tabelah.

Tabela 6.1-23: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo različice rotacijskega regresijskega gozda z agresivno metodo za ohranjanje variance s preostalimi najboljšimi metodami pri vseh deležih manjkajočih vrednosti (N=98)

1. metoda	2. metoda	Z	Signifikanca
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-0,331[a]	0,740
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-8,595[b]	0,000
Modelno drevo	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-8,227[b]	0,000
Rotacijski regresijski gozd (osnovna različica)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-8,508[a]	0,000
Rotacijski regresijski gozd (osnovna različica)	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-8,595[a]	0,000

[a] na podlagi negativnih rangov
[b] na podlagi pozitivnih rangov

Ob upoštevanju pogoja $z < -1{,}96$ pri $p < 0{,}05$ je razvidno, da se rotacijski regresijski gozd z agresivno metodo za ohranjanje variance signifikantno razlikuje od različice z neagresivno metodo za ohranjanje variance in modelnega drevesa, kar je bila 4. najbolje rangirana metoda. Če pri analizi upoštevamo vse deleže manjkajočih vrednosti, tega ne moremo trditi tudi za metodo orodja za statistično analizo SPSS, ki ni občutno slabša ali boljša od rotacijskega regresijskega gozda z agresivno metodo za ohranjanje variance.

Tabela 6.1-24: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo različice rotacijskega regresijskega gozda z agresivno metodo za ohranjanje variance s preostalimi najboljšimi metodami pri 25% in 50% manjkajočih vrednosti (N=28)

1. metoda	2. metoda	Z	Signifikanca
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-2,573[a]	0,010
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-4,623[b]	0,000
Modelno drevo	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-4,205[b]	0,000

[a] na podlagi negativnih rangov
[b] na podlagi pozitivnih rangov

Drugače se izkaže, kadar v test vključimo samo rezultate na podatkovnih množicah z več kot 20% manjkajočih vrednosti. Takrat lahko sklepamo, da metoda orodja za statistično analizo SPSS signifikantno bolje ohranja varianco kot preostale metode, tudi bolje kot rotacijski regresijski gozd z agresivno metodo za ohranjanje variance. Le-ta pa se še vedno izkaže kot občutno uspešnejši od preostalih metod.

Pri analizi moramo upoštevati, da smo metode na posameznih podatkovnih množicah ocenjevali s pomočjo rangov Friedmanovega testa, torej smo posamezne proporcionalne ocene ohranjanja variance na posamičnih atributih preslikali v intervalne vrednosti. To smo lahko storili, ker tako Friedmanov kot Wilcoxonov test operirata z rangi, pri čemer ju ne zanima dejanska razlike med sosednje rangiranima vrednostma. Če želimo še bolj natančno ovrednotiti razlike med metodami, si lahko pomagamo z grafi (npr. Graf 6.1-17). Na podlagi takšne analize lahko ugotovimo, da so razlike med posameznimi metodami lahko zelo majhne, še posebej pri majhnem odstotku manjkajočih vrednosti, hkrati pa so lahko precejšnje, kadar je odstotek manjkajočih vrednosti visok. Izmed vseh preverjenih metod pri tem najbolj izstopa metoda orodja za statistično analizo SPSS, ki se posveča predvsem ohranjanju variance in povprečne vrednosti, medtem ko je njena natančnost med najslabšimi in je primerljiva z natančnostjo dveh najbolj preprostih metod, ki manjkajoče vrednosti nadomeščata kar z ničlami oziroma s povprečnimi vrednostmi.

Čeprav lahko na podlagi opravljenih statističnih testov sklepamo, da obe različici rotacijskega regresijskega gozda, ki za ohranjanje variance uporabljata dodatno stohastično metodo, občutno bolje ohranjata varianco, kot osnovna varianta rotacijskega regresijskega gozda, smo za neposredno potrditev 3. hipoteze naše raziskave opravili dodatna Wilcoxonova testa. Rezultati se nahajajo v spodnji tabeli.

Tabela 6.1-25: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo osnovne različice rotacijskega regresijskega gozda z različicama z dodatnima metodama za ohranjanje variance (N=98)

1. metoda	2. metoda	Z	Signifikanca
Rotacijski regresijski gozd (osnovna različica)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-8,508[a]	0,000
Rotacijski regresijski gozd (osnovna različica)	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-8,595[a]	0,000

[a)] **na podlagi negativnih rangov**

Kot je bilo pričakovano, lahko zavrnemo ničelni hipotezi obeh Wilcoxonovih testov.

6.2 Učinkovitost metod pri mehanizmih MAR in NMAR

Kadar poskušamo z metodami strojnega učenja napovedati numerično vrednost na podlagi znanja, pridobljenega skozi učni proces, domnevamo, da med napovedano vrednostjo in preostalimi (znanimi) vrednostmi obstaja določena relacija oziroma odvisnost. Mehanizma nastanka naključnih (MAR) in nenaključnih (NMAR) manjkajočih vrednosti sicer nista nujno pogojena z obstojem takšne odvisnosti, vendar se posledice razlik med mehanizmi nastanka manjkajočih vrednosti pokažejo v celoti šele takrat, ko so vrednosti razrednega atributa odvisne od vrednosti enega ali več drugih atributov pripadajočega vzorca v podatkovni množici. Zaradi tega se pri primerjavi metod za nadomeščanje manjkajočih vrednosti, nastalih po mehanizmih MAR in NMAR, nismo zanašali na podatkovne množice, na katerih smo izvajali poskuse z manjkajočimi vrednostmi tipa MCAR. Da smo namreč lahko zagotovili zahtevano odvisnost med atributi, smo ustvarili lastno podatkovno množico s tremi neodvisnimi atributi in odvisnim, razrednim atributom. Vrednosti razrednega atributa te množice smo izračunali na podlagi poljubno generiranih vrednosti preostalih 3 atributov vsakega posameznega vzorca, in sicer kot rezultat multivariatnega polinoma tretje stopnje z naključno določenimi koeficienti, kot je bilo opisano v 5. poglavju.

Za manjkajoče vrednosti tipa MAR je značilno, da je verjetnost, da bo posamezna vrednost odvisnega atributa manjkala, odvisna od vrednosti nekega drugega atributa pripadajočega vzorca. To smo simulirali tako, da smo vzorce podatkovne množice razvrstili glede na velikost vrednosti izbranega neodvisnega atributa in nato odstranili 1%, 5%, 10%, 15%, 20%, 25% in 50% vseh vrednosti razrednega (odvisnega) atributa, pri čemer je verjetnost brisanja vrednosti odvisnega atributa posameznega vzorca eksponentno rasla z rastjo zaporedne številke vzorca (Slika 6.2-1). Na ta način smo ustvarili 21 različnih primerov podatkovne množice z naključno vstavljenimi manjkajočimi vrednostmi (7 različnih deležev manjkajočih vrednosti odvisnega atributa pri vsaki izmed treh razvrstitev glede na vrednosti izbranega neodvisnega atributa). Na vsaki izmed tako ustvarjenih podatkovnih množic smo izvedli 5 primerjav metod za nadomeščanje manjkajočih vrednosti in na podlagi pridobljenih ocen napak izračunali povprečne ocene, kot smo to naredili že pri manjkajočih vrednostih tipa MCAR. Rezultati so prikazani v spodnjih tabelah in grafih.

Iz grafov lahko razberemo, da se poleg linearne regresije vse tri različice našega rotacijskega regresijskega gozda izkažejo kot najbolj natančne metode, pri čemer ni opaziti, da bi pri varianti z agresivno metodo ohranjanja variance natančnost metode padla, pravzaprav prav nasprotno. Prav tako

lahko ugotovimo, da se z večanjem odstotka manjkajočih vrednosti natančnost napovedovanja opazno zmanjšuje le pri razvrstitvi po vrednostih tretjega atributa. To lahko pripišemo dejstvu, da tretji atribut predstavlja člen polinoma prve stopnje in so zaradi izbire vrednosti neodvisnih spremenljivk polinoma iz intervala [0,1] vrednosti tega atributa v povprečju večje kot vrednosti preostalih dveh neodvisnih atributov, ki predstavljata člena polinoma 2. in 3. stopnje. Zaradi tega je vpliv tretjega atributa na vrednosti razrednega atributa največji.

VHOD: odstotek manjkajočih vrednosti ***o***, podatkovna množica ***M***,
neodvisni atribut ***a***, odvisni atribut ***b***

PREŠTEJ število vzorcev ***n*** v podatkovni množici ***M***

IZRAČUNAJ število vseh manjkajočih vrednosti ***mv***:

- ***mv*** = ***n*** * ***o*** / 100

RAZVRSTI vzorce podatkovne množice ***M*** glede na vrednosti atributa ***a*** (od največje proti najmanjši)

PONAVLJAJ:

- naključno IZBERI celo število ***r*** iz intervala [0, n^4]
- ***z*** = $r^{1/4}$ (zaokroženo navzgor)
- IZBERI vzorec ***v*** z zaporedno številko ***z***
- ***d*** = 1
- ČE je ***v.b*** manjkajoča vrednost PONAVLJAJ:
    - ČE je (***z*** + ***d*** ≤ ***n***) IN (***z*** - ***d*** > 0) POTEM:
        - ***z*** = ***z*** + ***d***
        - IZBERI vzorec ***v*** z zaporedno številko ***z***
        - ČE je ***v.b*** manjkajoča vrednost POTEM:
            - ***z*** = ***z*** – 2 * ***d***
            - IZBERI vzorec ***v*** z zaporedno številko ***z***
            - ***z*** = ***z*** + ***d***
        - ***d*** = ***d*** + 1
- DOKLER ***v.b*** ni manjkajoča vrednost
- IZBRIŠI ***v.b***

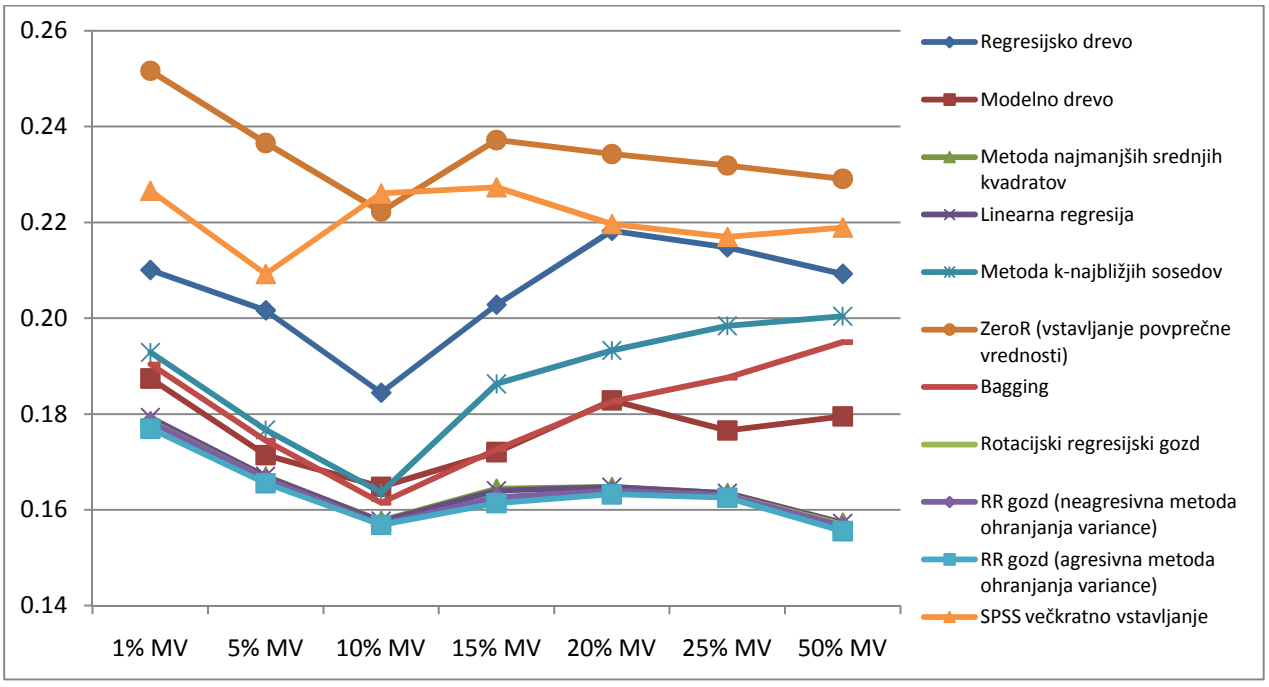
DOKLER število vstavljenih manjkajočih vrednosti ni enako ***mv***

Slika 6.2-1: Psevdokod algoritma za vstavljanje manjkajočih vrednosti po mehanizmu MAR

Tabela 6.2-1: Povprečne napake vstavljenih vrednosti razrednega atributa umetne podatkovne množice, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 1. neodvisnega atributa

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	2,10E-1 ± 3,14E-2	2,02E-1 ± 2,30E-2	1,84E-1 ± 6,13E-3	2,03E-1 ± 1,05E-2	2,18E-1 ± 2,71E-3	2,15E-1 ± 9,02E-3	2,09E-1 ± 5,87E-3
Modelno drevo	1,87E-1 ± 3,18E-2	1,71E-1 ± 1,17E-2	1,65E-1 ± 5,49E-3	1,72E-1 ± 2,64E-3	1,83E-1 ± 2,76E-3	1,77E-1 ± 1,26E-3	1,79E-1 ± 7,65E-3
MNSK	1,79E-1 ± 3,03E-2	1,67E-1 ± 9,42E-3	1,58E-1 ± 4,80E-3	1,64E-1 ± 8,19E-4	1,65E-1 ± 7,73E-4	1,64E-1 ± 3,52E-4	1,57E-1 ± 2,42E-4
Lin. Reg.	1,79E-1 ± 3,01E-2	1,67E-1 ± 9,44E-3	1,58E-1 ± 4,74E-3	1,64E-1 ± 7,65E-4	1,65E-1 ± 6,70E-4	1,63E-1 ± 2,13E-4	1,57E-1 ± 1,73E-4
K-NN	1,93E-1 ± 3,64E-2	1,77E-1 ± 3,98E-3	1,64E-1 ± 5,78E-3	1,86E-1 ± 1,89E-3	1,93E-1 ± 2,15E-3	1,98E-1 ± 3,15E-4	2,00E-1 ± 2,25E-4
ZeroR	2,52E-1 ± 2,65E-2	2,37E-1 ± 5,12E-3	2,22E-1 ± 4,89E-3	2,37E-1 ± 2,26E-3	2,34E-1 ± 1,77E-3	2,32E-1 ± 2,57E-4	2,29E-1 ± 2,13E-4
Vstavi 0	6,30E-1 ± 2,67E-2	6,06E-1 ± 8,65E-3	5,87E-1 ± 1,86E-3	5,92E-1 ± 2,85E-3	5,82E-1 ± 2,00E-3	5,71E-1 ± 5,61E-4	5,40E-1 ± 5,11E-4
Bagging	1,90E-1 ± 3,61E-2	1,74E-1 ± 1,32E-2	1,61E-1 ± 5,80E-3	1,73E-1 ± 5,37E-3	1,83E-1 ± 3,58E-3	1,88E-1 ± 1,38E-3	1,95E-1 ± 8,76E-4
Rot. reg. gozd	1,79E-1 ± 3,01E-2	1,66E-1 ± 9,60E-3	1,57E-1 ± 4,26E-3	1,62E-1 ± 4,73E-4	1,64E-1 ± 4,69E-4	1,63E-1 ± 2,87E-4	1,57E-1 ± 1,74E-4
RRG var. 1	1,78E-1 ± 3,02E-2	1,66E-1 ± 9,60E-3	1,57E-1 ± 4,27E-3	1,63E-1 ± 5,65E-4	1,64E-1 ± 4,24E-4	1,63E-1 ± 3,46E-4	1,57E-1 ± 1,72E-4
RRG var. 2	1,77E-1 ± 3,02E-2	1,66E-1 ± 9,49E-3	1,57E-1 ± 4,42E-3	1,61E-1 ± 8,63E-4	1,63E-1 ± 3,62E-4	1,63E-1 ± 8,09E-4	1,56E-1 ± 4,29E-4
SPSS	2,27E-1 ± 3,31E-2	2,09E-1 ± 4,20E-3	2,26E-1 ± 7,74E-4	2,27E-1 ± 3,51E-3	2,20E-1 ± 1,62E-2	2,17E-1 ± 1,12E-3	2,19E-1 ± 6,46E-3

*) Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance

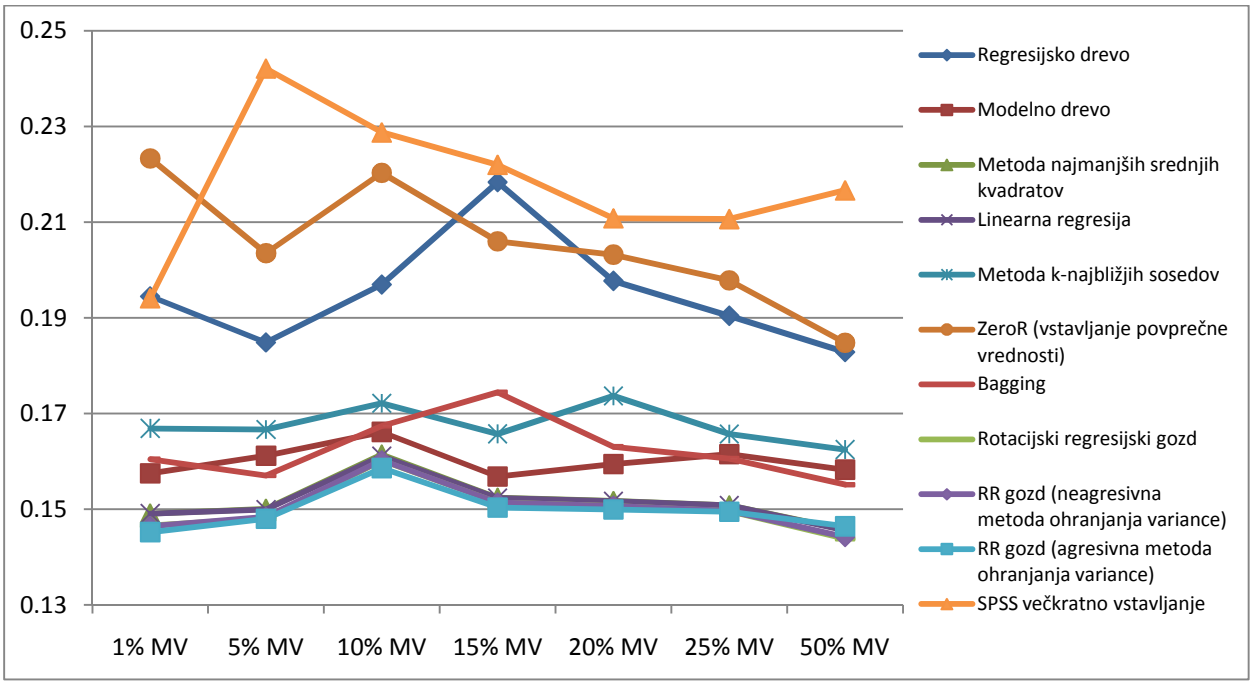


Graf 6.2-1: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 1. atributa (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.2-2: Povprečne napake vstavljenih vrednosti razrednega atributa umetne podatkovne množice, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 2. neodvisnega atributa

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	1,94E-1 ± 3,78E-2	1,85E-1 ± 4,70E-3	1,97E-1 ± 1,62E-2	2,18E-1 ± 1,29E-2	1,98E-1 ± 1,27E-2	1,90E-1 ± 5,25E-3	1,83E-1 ± 2,61E-4
Modelno drevo	1,58E-1 ± 2,41E-2	1,61E-1 ± 7,43E-3	1,66E-1 ± 5,35E-3	1,57E-1 ± 4,62E-3	1,59E-1 ± 4,39E-3	1,62E-1 ± 1,05E-3	1,58E-1 ± 2,32E-3
MNSK	1,49E-1 ± 2,64E-2	1,50E-1 ± 1,01E-2	1,61E-1 ± 3,06E-3	1,52E-1 ± 1,81E-3	1,52E-1 ± 5,81E-4	1,51E-1 ± 2,92E-4	1,46E-1 ± 7,39E-4
Lin. Reg.	1,49E-1 ± 2,62E-2	1,50E-1 ± 9,79E-3	1,61E-1 ± 2,76E-3	1,52E-1 ± 1,77E-3	1,52E-1 ± 6,69E-4	1,51E-1 ± 2,84E-4	1,46E-1 ± 4,54E-4
K-NN	1,67E-1 ± 3,31E-2	1,67E-1 ± 1,58E-2	1,72E-1 ± 2,32E-3	1,66E-1 ± 4,44E-3	1,74E-1 ± 4,13E-3	1,66E-1 ± 6,87E-4	1,62E-1 ± 2,86E-4
ZeroR	2,23E-1 ± 6,09E-2	2,04E-1 ± 1,22E-2	2,20E-1 ± 2,33E-3	2,06E-1 ± 1,37E-3	2,03E-1 ± 1,41E-3	1,98E-1 ± 6,00E-4	1,85E-1 ± 8,11E-5
Vstavi 0	5,56E-1 ± 4,57E-2	5,36E-1 ± 1,85E-2	5,47E-1 ± 7,86E-3	5,24E-1 ± 1,53E-3	5,18E-1 ± 1,23E-3	5,07E-1 ± 8,84E-4	4,87E-1 ± 8,24E-5
Bagging	1,60E-1 ± 2,51E-2	1,57E-1 ± 7,84E-3	1,67E-1 ± 5,48E-3	1,74E-1 ± 7,01E-3	1,63E-1 ± 1,74E-3	1,61E-1 ± 1,75E-3	1,55E-1 ± 5,03E-5
Rot. reg. gozd	1,47E-1 ± 2,62E-2	1,48E-1 ± 9,56E-3	1,60E-1 ± 2,57E-3	1,51E-1 ± 2,04E-3	1,51E-1 ± 6,05E-4	1,50E-1 ± 4,17E-4	1,44E-1 ± 5,64E-4
RRG var. 1	1,46E-1 ± 2,61E-2	1,48E-1 ± 9,39E-3	1,60E-1 ± 2,63E-3	1,51E-1 ± 2,06E-3	1,51E-1 ± 6,20E-4	1,50E-1 ± 4,51E-4	1,44E-1 ± 6,83E-4
RRG var. 2	1,45E-1 ± 2,62E-2	1,48E-1 ± 8,56E-3	1,59E-1 ± 2,77E-3	1,50E-1 ± 2,16E-3	1,50E-1 ± 5,75E-4	1,49E-1 ± 5,32E-4	1,46E-1 ± 6,51E-4
SPSS	1,94E-1 ± 8,11E-3	2,42E-1 ± 1,69E-2	2,29E-1 ± 1,09E-2	2,22E-1 ± 2,49E-3	2,11E-1 ± 1,25E-2	2,11E-1 ± 1,34E-3	2,17E-1 ± 5,42E-3

*) Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance

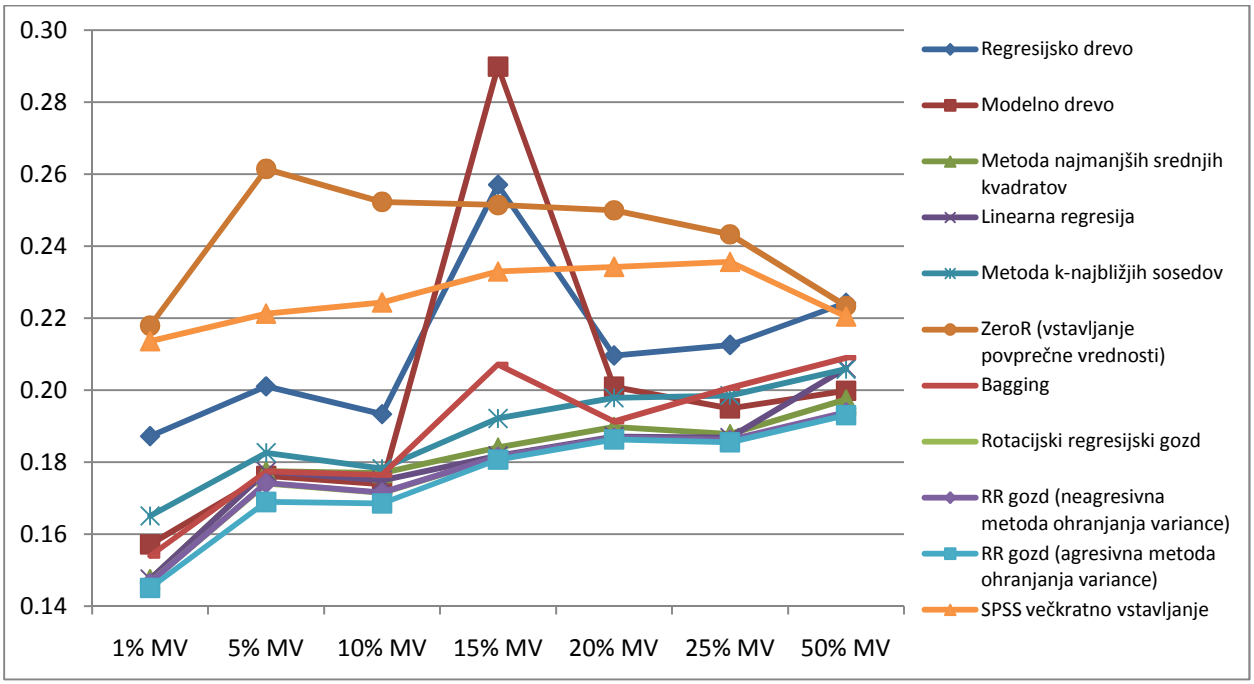


Graf 6.2-2: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 2. atributa (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.2-3: Povprečne napake vstavljenih vrednosti razrednega atributa umetne podatkovne množice, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 3. neodvisnega atributa

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	1,87E-1 ± 4,25E-2	2,01E-1 ± 1,78E-2	1,93E-1 ± 2,65E-2	2,57E-1 ± 2,01E-2	2,10E-1 ± 4,11E-3	2,13E-1 ± 6,16E-3	2,24E-1 ± 3,10E-4
Modelno drevo	1,57E-1 ± 2,24E-2	1,76E-1 ± 1,41E-2	1,74E-1 ± 7,32E-3	2,90E-1 ± 3,82E-3	2,01E-1 ± 5,84E-3	1,95E-1 ± 3,28E-3	2,00E-1 ± 1,42E-3
MNSK	1,47E-1 ± 2,28E-2	1,78E-1 ± 1,45E-2	1,77E-1 ± 2,49E-3	1,84E-1 ± 2,73E-3	1,90E-1 ± 1,45E-3	1,88E-1 ± 1,09E-3	1,97E-1 ± 9,94E-4
Lin. Reg.	1,47E-1 ± 2,27E-2	1,77E-1 ± 1,43E-2	1,75E-1 ± 2,31E-3	1,82E-1 ± 1,25E-3	1,87E-1 ± 1,69E-3	1,87E-1 ± 2,77E-4	2,06E-1 ± 1,47E-4
K-NN	1,65E-1 ± 3,64E-2	1,83E-1 ± 1,07E-2	1,78E-1 ± 3,70E-3	1,92E-1 ± 1,32E-3	1,98E-1 ± 1,96E-3	1,98E-1 ± 1,41E-3	2,06E-1 ± 9,09E-4
ZeroR	2,18E-1 ± 2,10E-2	2,61E-1 ± 1,70E-2	2,52E-1 ± 3,68E-3	2,51E-1 ± 1,71E-3	2,50E-1 ± 4,86E-4	2,43E-1 ± 2,63E-4	2,23E-1 ± 1,93E-4
Vstavi 0	6,00E-1 ± 2,19E-2	6,41E-1 ± 1,66E-2	6,27E-1 ± 4,59E-3	6,15E-1 ± 2,15E-3	6,02E-1 ± 9,11E-4	5,87E-1 ± 5,31E-4	5,32E-1 ± 2,41E-4
Bagging	1,54E-1 ± 2,86E-2	1,77E-1 ± 1,78E-2	1,76E-1 ± 9,39E-3	2,07E-1 ± 1,49E-2	1,91E-1 ± 3,39E-3	2,01E-1 ± 6,20E-3	2,09E-1 ± 4,40E-3
Rot. reg. gozd	1,47E-1 ± 2,35E-2	1,74E-1 ± 1,34E-2	1,71E-1 ± 5,42E-3	1,81E-1 ± 1,36E-3	1,87E-1 ± 1,87E-3	1,86E-1 ± 4,89E-4	1,93E-1 ± 9,13E-4
RRG var. 1	1,47E-1 ± 2,37E-2	1,74E-1 ± 1,34E-2	1,72E-1 ± 5,39E-3	1,81E-1 ± 1,35E-3	1,87E-1 ± 1,82E-3	1,86E-1 ± 4,50E-4	1,94E-1 ± 6,57E-4
RRG var. 2	1,45E-1 ± 2,53E-2	1,69E-1 ± 1,22E-2	1,69E-1 ± 6,83E-3	1,81E-1 ± 1,69E-3	1,86E-1 ± 1,91E-3	1,85E-1 ± 5,74E-4	1,93E-1 ± 1,19E-3
SPSS	2,14E-1 ± 1,93E-2	2,21E-1 ± 8,04E-3	2,24E-1 ± 2,01E-2	2,33E-1 ± 8,36E-4	2,34E-1 ± 1,08E-2	2,36E-1 ± 3,87E-3	2,20E-1 ± 1,48E-2

*) Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance



Graf 6.2-3: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 3. atributa (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Za lažje razločevanje razlik med ocenami natančnosti posameznih metod si lahko spet pomagamo z razvrstitvijo po rangih za vsak poskus posebej. Če pogledamo range v tabeli (Tabela 6.2-4), pričakovano opazimo, da je prepričljivo najslabša metoda vstavljanja ničel, ki ji sledi metoda vstavljanja povprečne vrednosti. Slabo se je odrezala tudi metoda orodja SPSS. Na drugi strani imamo vse tri različice rotacijskega regresijskega gozda, kjer se je na prvo mesto (z izjemo enega poskusa) vedno uvrstila varianta z agresivnim načinom ohranjanja variance. Preostali dve različici sta po rangih približno enakovredni in se v nobenem primeru nista uvrstili slabše kot na tretje mesto. Seveda nam rangiranje metod ne pove vsega o dejanskih razlikah v natančnosti. Dober primer tega je linearna regresija, ki se je v povprečju uvrščala slabše kot na 4. mesto, a lahko na podlagi grafov in tabel ugotovimo, da za rotacijskim regresijskim gozdom zaostaja zelo malo.

Tabela 6.2-4: Razvrstitev metod po rangih glede na natančnost pri posameznih poskusih na umetni podatkovni množici

Atrib. (% mv)	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
1. atr. (1%)	9	6	4	5	8	11	12	7	3	2	1	10
1. atr. (5%)	9	6	4	5	8	11	12	7	2	3	1	10
1. atr. (10%)	9	8	5	4	7	10	12	6	2	3	1	11
1. atr. (15%)	9	6	5	4	8	11	12	7	2	3	1	10
1. atr. (20%)	9	7	5	4	8	11	12	6	2	3	1	10
1. atr. (25%)	9	6	5	4	8	11	12	7	3	2	1	10
1. atr. (50%)	9	6	5	4	8	11	12	7	3	2	1	10
2. atr. (1%)	10	6	5	4	8	11	12	7	3	2	1	9
2. atr. (5%)	9	7	5	4	8	10	12	6	2	3	1	11
2. atr. (10%)	9	6	5	4	8	10	12	7	2	3	1	11
2. atr. (15%)	10	6	5	4	7	9	12	8	3	2	1	11
2. atr. (20%)	9	6	5	4	8	10	12	7	2	3	1	11
2. atr. (25%)	9	7	5	4	8	10	12	6	2	3	1	11
2. atr. (50%)	9	7	4	3	8	10	12	6	1	2	5	11
3. atr. (1%)	9	7	4	5	8	11	12	6	3	2	1	10
3. atr. (5%)	9	4	7	5	8	11	12	6	2	3	1	10
3. atr. (10%)	9	4	7	5	8	11	12	6	2	3	1	10
3. atr. (15%)	10	11	5	4	6	9	12	7	2	3	1	8
3. atr. (20%)	9	8	5	4	7	11	12	6	2	3	1	10
3. atr. (25%)	9	6	5	4	7	11	12	8	2	3	1	10
3. atr. (50%)	11	5	4	7	6	10	12	8	2	3	1	9
Povprečni rang	**9,24** ± 0,54	**6,43** ± 1,47	**4,95** ± 0,80	**4,33** ± 0,80	**7,62** ± 0,67	**10,48** ± 0,68	**12,00** ± 0,00	**6,71** ± 0,72	**2,24** ± 0,54	**2,67** ± 0,48	**1,19** ± 0,87	**10,14** ± 0,79

Tudi tokrat smo za preverjanje ničelne hipoteze o enakovrednosti posameznih metod uporabili Friedmanov test (Tabela 6.2-5).

Tabela 6.2-5: Friedmanov test za primerjavo metod po natančnosti napovedovanja manjkajočih vrednosti razrednega atributa

N	χ^2	stopnje prostosti	p
21	220,128	11	0,000

Za potrditev naše domneve, da so variante rotacijskega regresijskega gozda signifikantno bolj natančne od preostalih metod, smo uporabili Wilcoxonov test predznačenih rangov. V primerjavo smo vključili najboljših 6 metod na podlagi Friedmanovih povprečnih rangov.

Tabela 6.2-6: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo različic rotacijskega regresijskega gozda med seboj in s preostalimi najboljšimi metodami pri mehanizmu MAR (N=21)

1. metoda	2. metoda	Z	Signifikanca
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-1,894[a]	0,058
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-3,354[b]	0,001
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-3,354[b]	0,001
Rotacijski regresijski gozd	Modelno drevo	-4,015[b]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Modelno drevo	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Modelno drevo	-4,015[b]	0,000
Rotacijski regresijski gozd	Metoda najmanjših srednjih kvadratov	-4,015[b]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-3,980[b]	0,000
Rotacijski regresijski gozd	Linearna regresija	-4,015[b]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Linearna regresija	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Linearna regresija	-3,910[b]	0,000

[a] na podlagi negativnih rangov
[b] na podlagi pozitivnih rangov

Če upoštevamo pogoj $z < -1,96$ pri $p < 0,05$, ne moremo trditi, da je različica rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance signifikantno manj natančna pri napovedovanju manjkajočih vrednosti kot rotacijski regresijski gozd brez dodatne metode za ohranjanje variance, čeprav se je v kar 15 od 21 poskusov izkazala slabše. Vse druge primerjane metode so se izkazale za signifikantno različne, pri čemer je Wilcoxonov test pokazal, da so vse tri različice rotacijskega

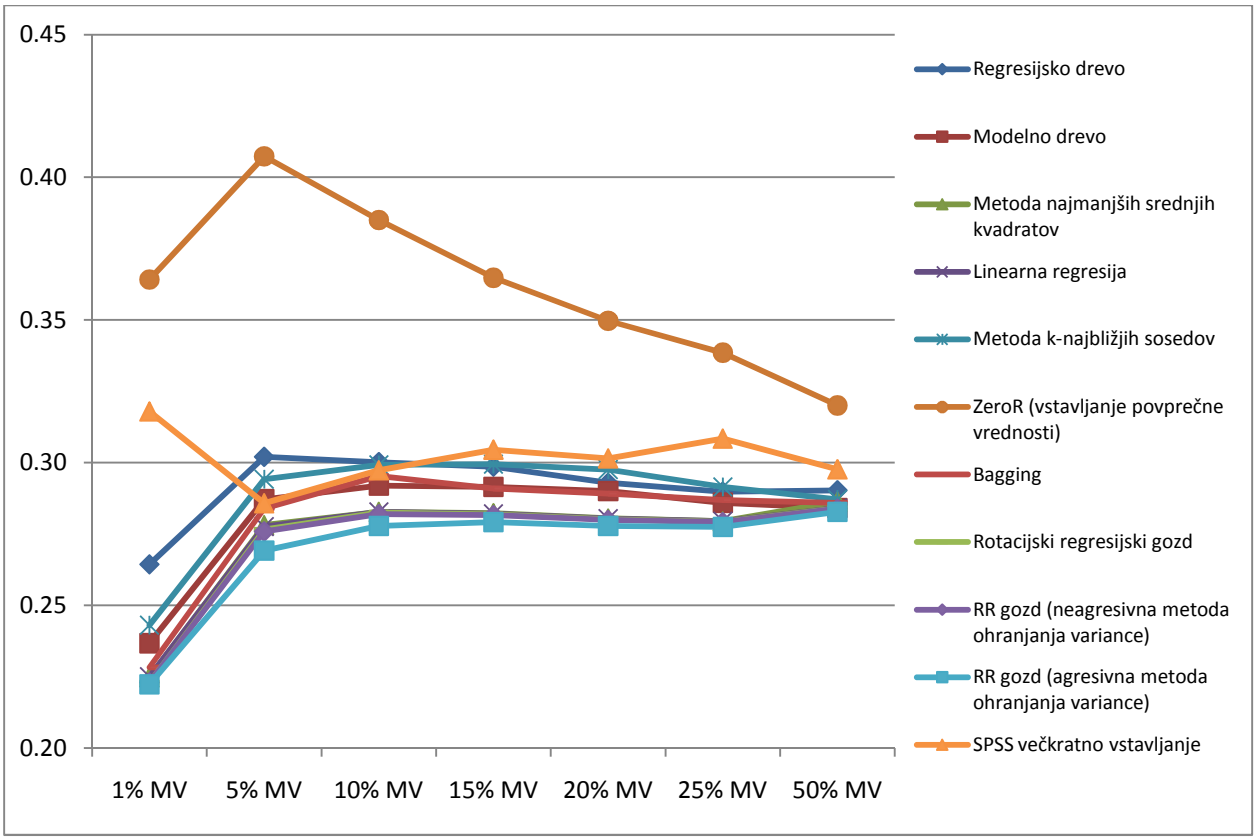
regresijskega gozda natančnejše od preostalih najboljših metod (modelno drevo, linearna regresija in metoda najmanjših srednjih kvadratov).

Kakor v primeru manjkajočih vrednosti tipa MAR smo tudi simulacijo mehanizma nastanka nenaključno manjkajočih vrednosti (NMAR) izvedli na umetni podatkovni množici. Za manjkajoče vrednosti tipa NMAR je namreč značilno, da je verjetnost, da bo posamezna vrednost izbranega atributa manjkala, odvisna kar od te vrednosti. Za vstavljanje različnih deležev manjkajočih vrednosti v našo umetno podatkovno množico smo uporabili isti postopek kot pri manjkajočih vrednostih tipa MAR, le da smo vzorce podatkovne množice razvrstili glede na velikost vrednosti razrednega atributa in nato odstranili 1%, 5%, 10%, 15%, 20%, 25% in 50% vseh vrednosti tega atributa, pri čemer je verjetnost brisanja vrednosti posameznega vzorca eksponentno rasla z rastjo zaporedne številke vzorca (Slika 6.2-1). Za vsakega izmed 7 različnih deležev manjkajočih vrednosti odvisnega atributa smo izvedli 5 primerjav metod za nadomeščanje manjkajočih vrednosti in na podlagi pridobljenih ocen napak izračunali povprečne ocene, kot smo to naredili že pri prejšnjih mehanizmih. Rezultati so prikazani v spodnji tabeli in grafu.

Tabela 6.2-7: Povprečne napake vstavljenih vrednosti razrednega atributa umetne podatkovne množice, kadar je mehanizem nastanka manjkajočih vrednosti NMAR

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	2.64E-1 ± 6.53E-2	3.02E-1 ± 2.08E-2	3.00E-1 ± 6.17E-3	2.99E-1 ± 5.72E-3	2.93E-1 ± 2.23E-3	2.90E-1 ± 2.93E-3	2.90E-1 ± 2.62E-3
Modelno drevo	2.37E-1 ± 3.32E-2	2.87E-1 ± 4.37E-3	2.92E-1 ± 2.03E-3	2.92E-1 ± 2.53E-3	2.90E-1 ± 1.01E-3	2.86E-1 ± 2.89E-4	2.84E-1 ± 9.53E-4
MNSK	2.25E-1 ± 2.62E-2	2.78E-1 ± 5.22E-3	2.83E-1 ± 2.78E-3	2.82E-1 ± 1.46E-3	2.80E-1 ± 1.28E-4	2.80E-1 ± 9.08E-5	2.86E-1 ± 6.78E-5
Lin. Reg.	2.25E-1 ± 2.59E-2	2.78E-1 ± 4.91E-3	2.83E-1 ± 2.63E-3	2.82E-1 ± 1.12E-3	2.80E-1 ± 7.04E-4	2.80E-1 ± 9.08E-5	2.84E-1 ± 8.38E-5
K-NN	2.43E-1 ± 3.46E-2	2.94E-1 ± 2.67E-3	2.99E-1 ± 4.53E-3	2.99E-1 ± 1.90E-3	2.98E-1 ± 4.82E-4	2.92E-1 ± 2.54E-4	2.87E-1 ± 1.54E-4
ZeroR	3.64E-1 ± 4.57E-2	4.07E-1 ± 8.60E-3	3.85E-1 ± 1.53E-3	3.65E-1 ± 8.47E-4	3.50E-1 ± 4.29E-4	3.39E-1 ± 1.42E-4	3.20E-1 ± 7.24E-5
Vstavi 0	7.93E-1 ± 4.49E-2	8.19E-1 ± 7.79E-3	7.79E-1 ± 1.60E-3	7.43E-1 ± 9.02E-4	7.13E-1 ± 4.57E-4	6.87E-1 ± 7.86E-5	5.98E-1 ± 3.86E-5
Bagging	2.28E-1 ± 2.46E-2	2.84E-1 ± 5.91E-3	2.95E-1 ± 4.18E-3	2.91E-1 ± 3.49E-3	2.89E-1 ± 3.67E-3	2.87E-1 ± 1.05E-3	2.86E-1 ± 8.35E-4
Rot. reg. gozd	2.23E-1 ± 2.57E-2	2.77E-1 ± 5.68E-3	2.83E-1 ± 2.58E-3	2.82E-1 ± 1.18E-3	2.80E-1 ± 6.33E-4	2.79E-1 ± 5.82E-5	2.84E-1 ± 6.36E-5
RRG var. 1	2.23E-1 ± 2.57E-2	2.76E-1 ± 5.84E-3	2.82E-1 ± 2.65E-3	2.82E-1 ± 1.20E-3	2.80E-1 ± 7.86E-4	2.79E-1 ± 7.44E-5	2.84E-1 ± 6.95E-5
RRG var. 2	2.22E-1 ± 2.69E-2	2.69E-1 ± 8.43E-3	2.78E-1 ± 3.64E-3	2.79E-1 ± 2.30E-3	2.78E-1 ± 1.67E-3	2.78E-1 ± 3.37E-4	2.83E-1 ± 4.27E-4
SPSS	3.18E-1 ± 4.38E-2	2.86E-1 ± 3.16E-2	2.97E-1 ± 1.77E-2	3.05E-1 ± 1.27E-2	3.01E-1 ± 1.25E-3	3.08E-1 ± 1.17E-2	2.98E-1 ± 2.74E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**



Graf 6.2-4: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je mehanizem nastanka manjkajočih vrednosti NMAR (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Ker imamo v tem primeru na voljo samo 7 različnih primerjav na podlagi povprečnih ocen, smo pri rangiranju metod s pomočjo Friedmanovega testa uporabili ocene napak vseh 5 ponovitev poskusov in tako dobili 35 neodvisnih primerjav. Povprečni rangi in rezultati Friedmanovega testa se nahajajo v tabelah (Tabela 6.2-8, Tabela 6.2-9).

Tabela 6.2-8: Friedmanova razvrstitev metod po rangih glede natančnosti napovedovanja manjkajočih vrednosti tipa NMAR v umetni podatkovni množici

	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Povprečni rang*	8,43	6,43	5,03	4,63	8,63	11,00	12,00	6,43	3,09	2,14	1,23	8,97

*) Povprečni rang po Friedmanovem testu (N=35, χ^2=348,539, sp=11, sig.=0,000)

Tabela 6.2-9: Friedmanov test za primerjavo metod po natančnosti napovedovanja manjkajočih vrednosti tipa NMAR v umetni podatkovni množici

N	χ^2	stopnje prostosti	p
35	348,539	11	0,000

Na podlagi povprečnih rangov lahko domnevamo, da so vse tri različice rotacijskega regresijskega gozda natančnejše od preostalih metod. Domnevo smo preverili z Wilcoxonovim testom, pri čemer smo v primerjavo vključili 7 najbolje rangiranih metod.

Tabela 6.2-10: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo različic rotacijskega regresijskega gozda med seboj in s preostalimi najboljšimi metodami pri mehanizmu NMAR (N=35)

1. metoda	2. metoda	Z	Signifikanca
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-4,860[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-4,860[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-4,860[a]	0,000
Rotacijski regresijski gozd	Modelno drevo	-5,160[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Modelno drevo	-5,160[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Modelno drevo	-5,160[a]	0,000
Rotacijski regresijski gozd	Metoda najmanjših srednjih kvadratov	-4,979[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-5,160[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-5,160[a]	0,000
Rotacijski regresijski gozd	Linearna regresija	-5,160[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Linearna regresija	-5,160[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Linearna regresija	-5,160[a]	0,000
Rotacijski regresijski gozd	Bagging	-4,865[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Bagging	-4,865[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Bagging	-4,947[a]	0,000

[a] na podlagi negativnih rangov

Iz tabele lahko razberemo, da so pri $p < 0,05$ vse tri različice rotacijskega regresijskega gozda signifikantno bolj natančne od naslednjih štirih najbolje rangiranih metod (linearna regresija, metoda najmanjših srednjih kvadratov, bagging in modelno drevo). Obenem lahko trdimo, da se je varianta rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance izkazala za signifikantno uspešnejšo od različice z neagresivno metodo ohranjanja variance, ki pa je spet signifikantno boljša od osnovnega rotacijskega regresijskega gozda.

Na podlagi tabele (Tabela 6.2-7) in grafa (Graf 6.2-4) opazimo, da se metode med seboj bistveno ne razlikujejo. Če iz primerjave izključimo dve najslabši metodi (vstavljanje ničel in vstavljanje povprečne

vrednosti), se razlike med natančnostjo najslabše in najboljše metode gibljejo od Δ_{TRMSE} = 0,015 (pri 50% manjkajočih vrednosti) do Δ_{TRMSE} = 0,096 (pri 1% manjkajočih vrednosti), hkrati pa so ocene napak teh metod relativno visoke (0,280 ± 0,021), če upoštevamo, da je standardni odklon vrednosti izbranega atributa enak σ_a= 0,184.

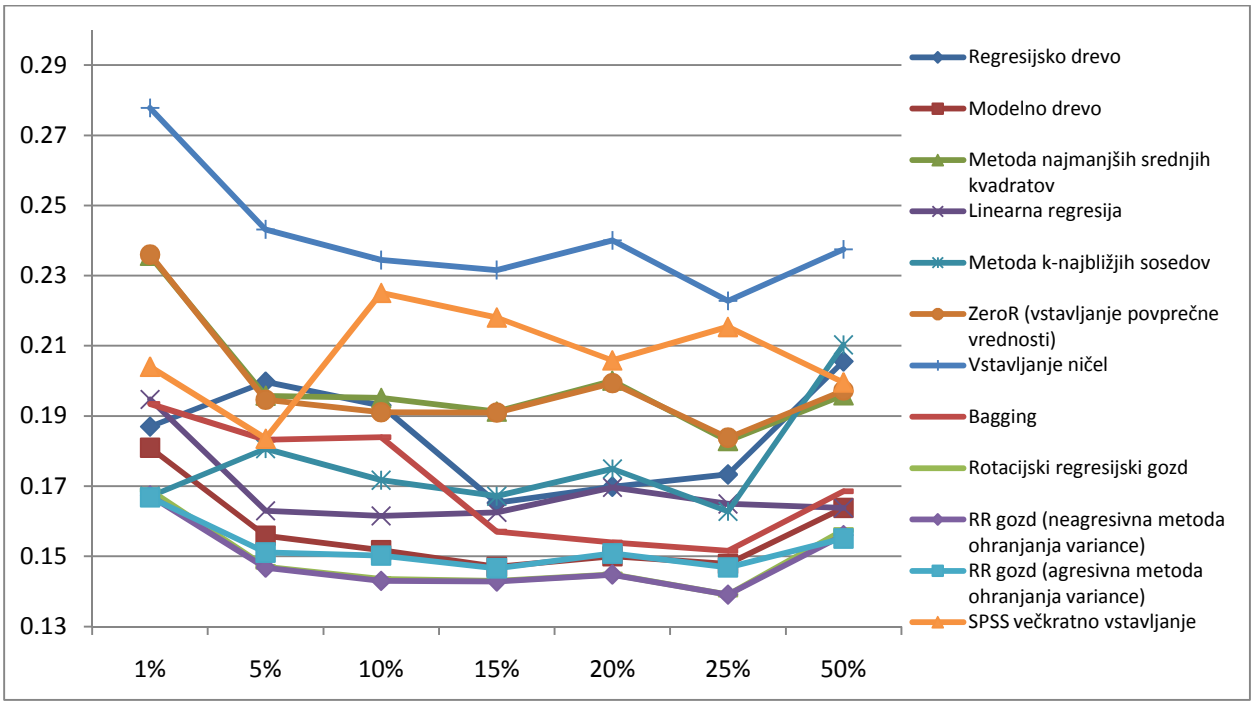
Če izračunamo povprečni koeficient korelacije (glej formulo 2.7) za izbrane metode pri 50% manjkajočih vrednosti, dobimo $\overline{CC}$ = 0,462 ± 0,058, torej skoraj na sredi med popolno neodvisnostjo in popolno odvisnostjo napovedanih in dejanskih vrednosti. Seveda bi bili rezultati vseh metod, ki temeljijo na linearni regresiji, veliko boljši, če bi pri zagotavljanju odvisnosti razrednega atributa umetne podatkovne množice namesto polinomske funkcije uporabili linearno, vendar bi bila ocena, temelječa na takšni množici vzorcev, nerealna. Koeficienti korelacije, ki smo jih izračunali pri napovedovanju manjkajočih vrednosti tipa MCAR na javno dostopnih podatkovnih množicah, se od atributa do atributa močno razlikujejo – od takšnih, ki pričajo o skoraj popolni soodvisnosti, do celo negativnih. Povprečni koeficient korelacije, ki smo ga izračunali na poskusih na umetni podatkovni množici se tako zdi dober kompromis med obema skrajnostma. Za primerjavo si oglejmo še, kako se metode izkažejo na realni podatkovni množici.

Izmed 14 podatkovnih množic in več kot 400 atributov smo izbrali 8. atribut podatkovne množice *Concrete*, pri katerem smo med preverjanjem metod na manjkajočih vrednostih tipa MCAR izračunali povprečen koeficient korelacije $\overline{CC}$ = 0,466 ± 0,137, torej primerljiv s tistim na umetni podatkovni množici. Povprečne napake metod pri različnih stopnjah manjkajočih vrednosti se nahajajo v spodnji tabeli in grafu. Opazimo lahko, da se napake posameznih metod precej bolj razlikujejo, kot pri poskusih na umetni podatkovni množici. Tudi če iz primerjave izključimo kar 4 najslabše metode (vstavljanje povprečnih vrednosti, vstavljanje ničel, metoda najmanjših srednjih kvadratov, metoda orodja SPSS), se razlike med natančnostjo najslabše in najboljše metode gibljejo od Δ_{TRMSE} = 0,0243 (pri 15% manjkajočih vrednosti) do Δ_{TRMSE} = 0,552 (pri 50% manjkajočih vrednosti), medtem ko so ocene napak teh metod relativno nizke (0,163 ± 0,018), saj se nahajajo znotraj standardnega odklona vrednosti izbranega atributa σ_a= 0,174.

Tabela 6.2-11: Povprečne napake vstavljenih vrednosti 8. atributa podatkovne množice *Concrete*, kadar je mehanizem nastanka manjkajočih vrednosti NMAR

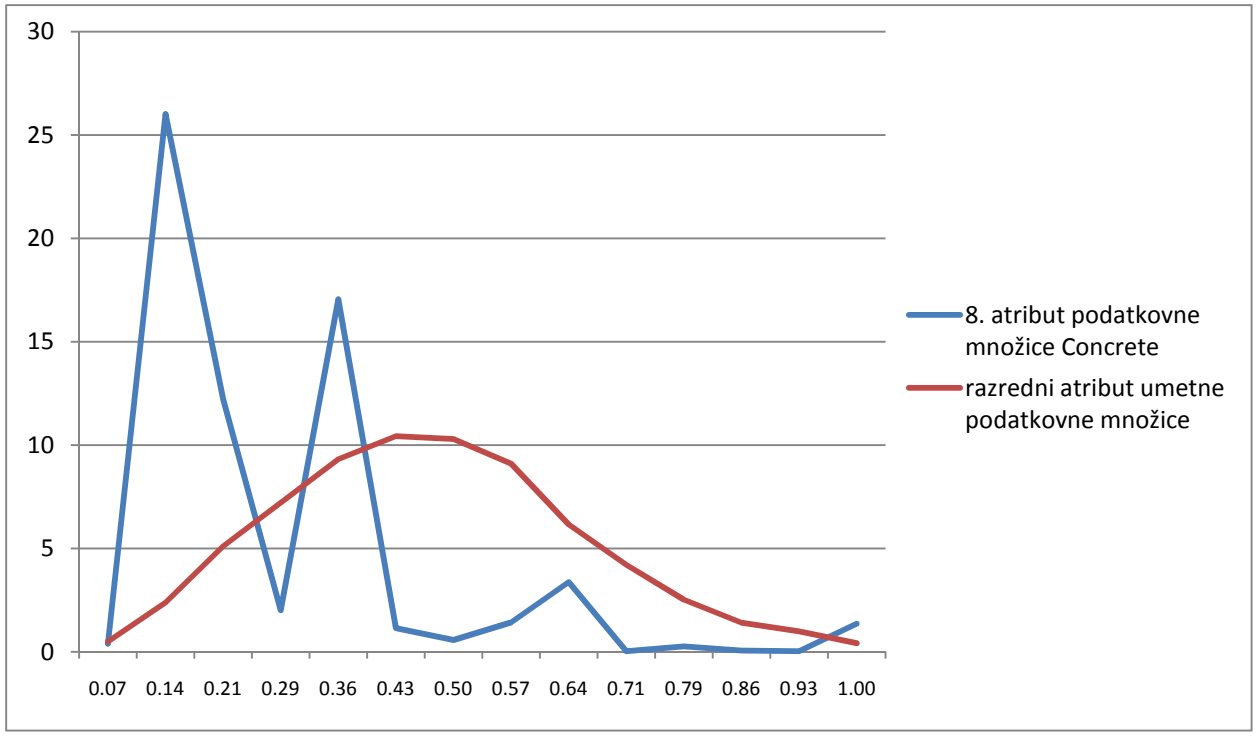
Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	1,87E-1 ± 4,50E-2	2,00E-1 ± 4,05E-2	1,93E-1 ± 1,48E-2	1,65E-1 ± 1,00E-2	1,70E-1 ± 1,66E-2	1,73E-1 ± 1,83E-2	2,06E-1 ± 2,47E-2
Modelno drevo	1,81E-1 ± 5,54E-2	1,56E-1 ± 2,88E-2	1,52E-1 ± 4,52E-3	1,47E-1 ± 3,62E-3	1,50E-1 ± 4,48E-4	1,48E-1 ± 4,03E-4	1,64E-1 ± 2,17E-4
MNSK	2,36E-1 ± 1,05E-1	1,96E-1 ± 3,80E-2	1,95E-1 ± 5,94E-3	1,91E-1 ± 6,22E-3	2,00E-1 ± 1,19E-3	1,83E-1 ± 3,04E-3	1,96E-1 ± 3,04E-4
Lin. Reg.	1,95E-1 ± 5,74E-2	1,63E-1 ± 2,31E-2	1,62E-1 ± 3,15E-3	1,63E-1 ± 4,83E-3	1,70E-1 ± 8,97E-4	1,65E-1 ± 1,17E-3	1,64E-1 ± 6,32E-4
K-NN	1,67E-1 ± 7,58E-2	1,81E-1 ± 3,11E-2	1,72E-1 ± 6,41E-3	1,67E-1 ± 7,86E-3	1,75E-1 ± 1,28E-4	1,63E-1 ± 1,26E-3	2,10E-1 ± 5,20E-4
ZeroR	2,36E-1 ± 1,05E-1	1,95E-1 ± 3,75E-2	1,91E-1 ± 5,66E-3	1,91E-1 ± 8,37E-3	1,99E-1 ± 4,32E-5	1,84E-1 ± 3,42E-3	1,97E-1 ± 2,12E-4
Vstavi 0	2,78E-1 ± 1,20E-1	2,43E-1 ± 3,99E-2	2,34E-1 ± 7,49E-3	2,32E-1 ± 8,85E-3	2,40E-1 ± 2,43E-5	2,23E-1 ± 3,68E-3	2,37E-1 ± 3,10E-4
Bagging	1,93E-1 ± 4,78E-2	1,83E-1 ± 2,67E-2	1,84E-1 ± 2,33E-2	1,57E-1 ± 1,23E-2	1,54E-1 ± 5,64E-3	1,52E-1 ± 8,80E-3	1,69E-1 ± 1,04E-3
Rot. reg. gozd	1,69E-1 ± 5,59E-2	1,47E-1 ± 3,02E-2	1,44E-1 ± 3,38E-3	1,43E-1 ± 5,00E-3	1,45E-1 ± 1,74E-3	1,39E-1 ± 4,07E-3	1,57E-1 ± 1,61E-3
RRG var. 1	1,67E-1 ± 5,55E-2	1,47E-1 ± 3,05E-2	1,43E-1 ± 3,83E-3	1,43E-1 ± 4,68E-3	1,45E-1 ± 1,15E-3	1,39E-1 ± 3,75E-3	1,56E-1 ± 2,57E-3
RRG var. 2	1,67E-1 ± 4,95E-2	1,51E-1 ± 2,55E-2	1,50E-1 ± 4,00E-3	1,47E-1 ± 2,63E-3	1,51E-1 ± 2,07E-3	1,47E-1 ± 4,05E-3	1,55E-1 ± 4,03E-3
SPSS	2,04E-1 ± 1,36E-2	1,84E-1 ± 4,09E-2	2,25E-1 ± 7,08E-4	2,18E-1 ± 1,28E-2	2,06E-1 ± 1,93E-2	2,15E-1 ± 2,06E-2	2,00E-1 ± 5,85E-3

*) Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance



Graf 6.2-5: Ocene povprečnih napak pri različnih stopnjah manjkajočih vrednosti 8. atributa podatkovne množice *Concrete*, kadar je mehanizem nastanka manjkajočih vrednosti NMAR

Tudi iz grafov (Graf 6.2-4 in Graf 6.2-5) lahko razberemo različen vzorec velikosti napake v odvisnosti od deleža manjkajočih vrednosti. Medtem ko se ocene napak pri poskusih na umetni podatkovni množici izrazito povečajo, ko delež manjkajočih vrednosti naraste z 1% na 5%, je pri podatkovni množici Concrete ravno obratno – povprečna ocena napake se pri večini metod zmanjša. Razlog teh razlik najdemo, če si ogledamo graf porazdelitve vrednosti razrednega atributa umetne podatkovne množice oz. 8. atributa podatkovne množice *Concrete* (Graf 6.2-6).



Graf 6.2-6: Frekvenca pojavitev posameznih intervalov vrednosti razrednega atributa umetne podatkovne množice in 8. atributa podatkovne množice *Concrete*

Opazimo, da je porazdelitev vrednosti razrednega atributa umetne podatkovne množice bolj ali manj normalna, medtem ko je porazdelitev atributa podatkovne množice *Concrete* izjemno asimetrična. Zaradi tega večanje odstotka manjkajočih ekstremnih vrednosti, ki jih je težko napovedati zaradi njihove majhne frekvence, ne privede do povečanja povprečne napake, ki se opazno spremeni šele pri 50% manjkajočih vrednosti.

Ne glede na prikazane razlike je videti, da so se tudi na izbranem primeru realne podatkovne množice najbolje izkazale različice rotacijskega regresijskega gozda. Za potrebe statističnih testov, s katerimi smo preverili to domnevo, smo tudi tokrat uporabili ocene napak vseh petih ponovitev poskusov (N=35). Rezultati Friemdanovega in Wilcoxonovih testov se nahajajo v tabelah (Tabela 6.2-12, Tabela 6.2-13 in Tabela 6.2-14).

Tabela 6.2-12: Friedmanova razvrstitev metod po rangih glede natančnosti napovedovanja manjkajočih vrednosti tipa NMAR v podatkovni množici *Concrete*

	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Povprečni rang*	7,57	4,39	9,00	5,99	7,09	9,14	11,74	6,09	2,23	2,09	3,09	9,60

*) Povprečni rang po Friedmanovem testu (N=35, χ^2=285,959, sp=11, sig.=0,000)

Tabela 6.2-13: Friedmanov test za primerjavo metod po natančnosti napovedovanja manjkajočih vrednosti tipa NMAR v podatkovni množici *Concrete*

N	χ^2	stopnje prostosti	p
35	285,959	11	0,000

Na podlagi povprečnih rangov lahko domnevamo, da so vse tri različice rotacijskega regresijskega gozda natančnejše od preostalih metod. Domnevo smo preverili z Wilcoxonovim testom, pri čemer smo v primerjavo vključili 6 najbolje rangiranih metod.

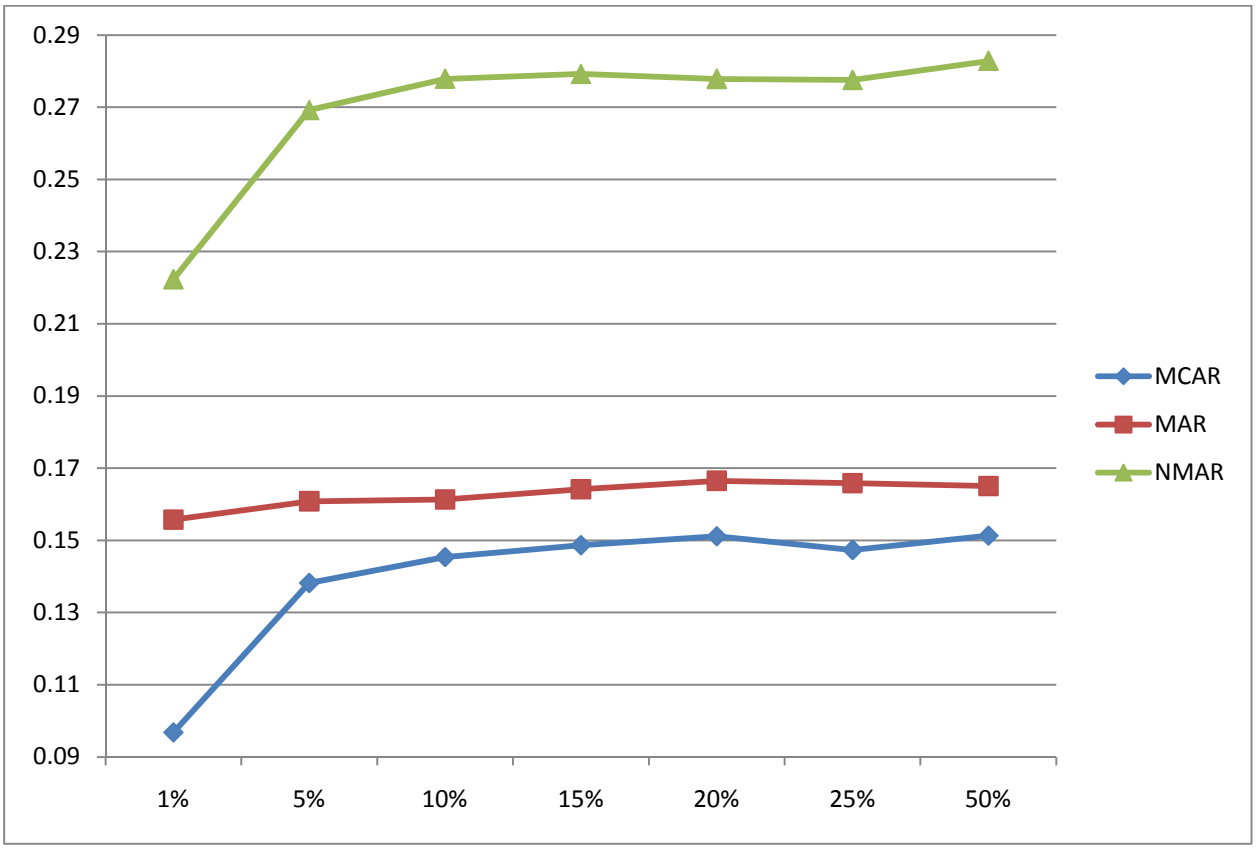
Tabela 6.2-14: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo različic rotacijskega regresijskega gozda med seboj in s preostalimi najboljšimi metodami pri mehanizmu NMAR na podatkovni množici *Concrete* (N=35)

1. metoda	2. metoda	Z	Signifikanca
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-2,060[a]	0,039
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-2,488[b]	0,013
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-2,847[b]	0,004
Rotacijski regresijski gozd	Modelno drevo	-5,159[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Modelno drevo	-5,143[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Modelno drevo	-2,997[a]	0,003
Rotacijski regresijski gozd	Linearna regresija	-5,159[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Linearna regresija	-5,159[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Linearna regresija	-4,799[a]	0,000
Rotacijski regresijski gozd	Bagging	-5,110[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Bagging	-5,078[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Bagging	-4,668[a]	0,000

[a] na podlagi negativnih rangov
[b] na podlagi pozitivnih rangov

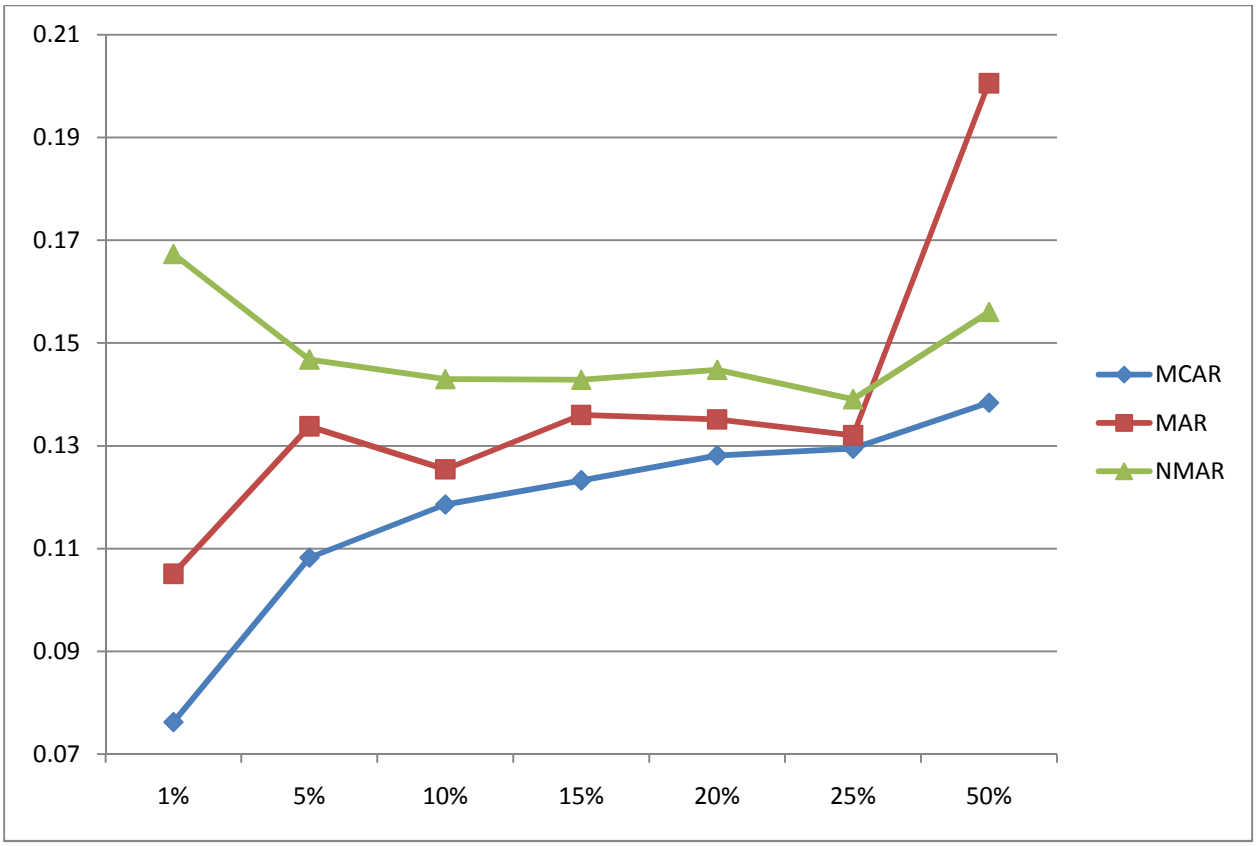
Ob upoštevanju pogoja $z < -1{,}96$ pri $p < 0{,}05$ je razvidno, da so vse tri različice rotacijskega regresijskega gozda signifikantno bolj natančne od naslednjih treh najbolje rangiranih metod (linearna regresija, bagging in modelno drevo). Za razliko od poskusov, opravljenih na umetni množici, kjer je bila najbolj natančna različica rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance, sta se tokrat preostali dve varianti naše metode izkazali za statistično uspešnejši.

Omenili smo že, da je napovedovanje manjkajočih vrednosti, nastalih po mehanizmu NMAR, najzahtevnejše. Oglejmo si, kako se s spremembo mehanizma nastanka manjkajočih vrednosti spreminjajo povprečne ocene dveh najuspešnejših metod: rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance na umetni podatkovni množici in rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance na podatkovni množici *Concrete* (Graf 6.2-7 in Graf 6.2-8).



Graf 6.2-7: Ocene povprečnih napak rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance ob napovedovanju različnih deležev manjkajočih vrednosti tipa MCAR, MAR in NMAR v umetni podatkovni množici

Povprečne napake nadomeščanja pri mehanizmu MAR so bile izračunane kot aritmetična povprečja napak, izračunanih na podlagi odvisnosti razrednega atributa od vrednosti vsakega izmed neodvisnih atributov (vseh preostalih).



Graf 6.2-8: Ocene povprečnih napak rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance ob napovedovanju različnih deležev manjkajočih vrednosti tipa MCAR, MAR in NMAR v podatkovni množici *Concrete* (8. atribut)

Opazimo, da so rezultati najboljši pri mehanizmu MCAR, tako pri eni kot pri drugi podatkovni množici. Pri tem moramo upoštevati tudi, da smo pri simulaciji mehanizma MCAR vstavljali manjkajoče vrednosti po celotni množici hkrati in ne samo znotraj izbranega atributa. Tako je bilo zaradi dodatnega redčenja informacij v učnih množicah delo metod za napovedovanje manjkajočih vrednosti še dodatno otežkočeno. Pričakujemo, da bi do podobnih rezultatov prišli na vseh množicah, na katerih smo izvajali poskuse, saj je izguba razpona učnih informacij hujša, kadar ni naključna. Drugo skrajnost predstavljajo manjkajoče vrednosti tipa NMAR, kjer vzrok njihovega nastanka leži v njih samih. Zaradi tega je pogosto zelo težko natančno nadomestiti manjkajoče vrednosti na podlagi vrednosti preostalih atributov. Rezultat tega so višje povprečne napake že pri majhnih odstotkih manjkajočih vrednosti.

6.2.1 Ohranjanje variance

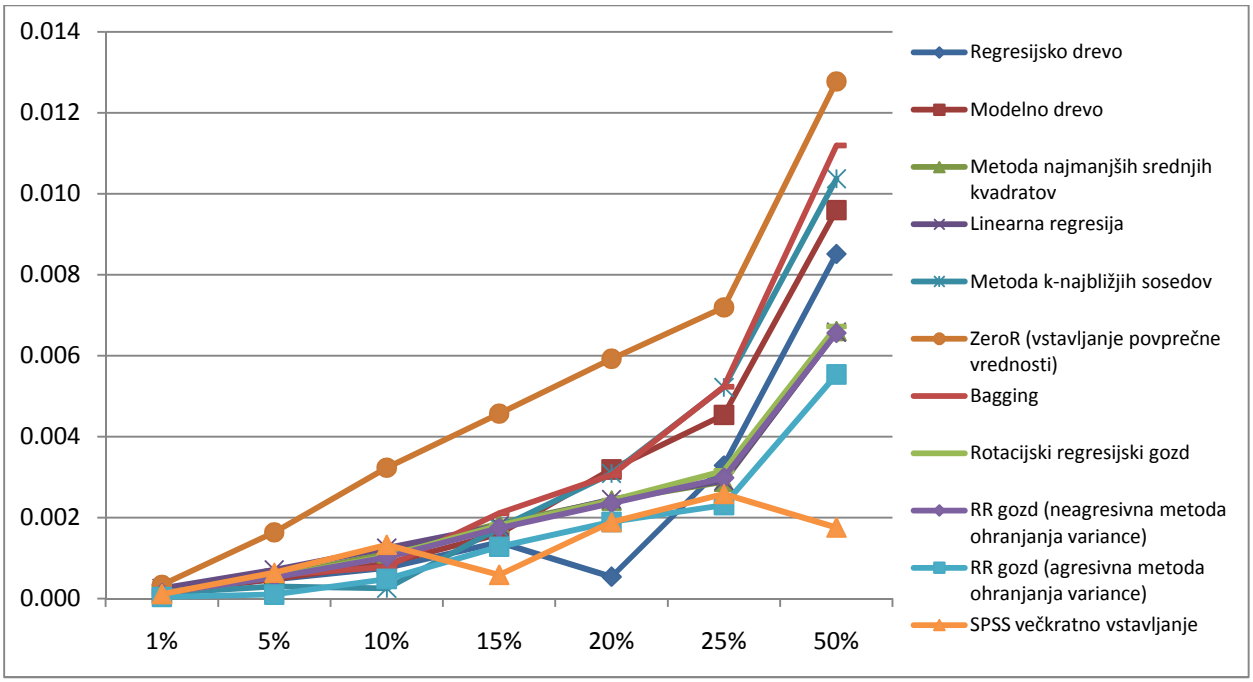
Ohranjanje variance pri nadomeščanju manjkajočih vrednosti tipa MAR in NMAR smo ocenjevali na enak način kot pri mehanizmu MCAR, le da tokrat nismo rangirali metod po posameznih poskusih, saj smo imeli opravka z majhnim številom atributov, ki smo jih lahko posebej obravnavali.

Pri nadomeščanju manjkajočih vrednosti tipa MAR smo uporabili že opisano umetno podatkovno množico s tremi neodvisnimi atributi. Ohranjanje variance pri nadomeščanju različnih odstotkov manjkajočih vrednosti razrednega atributa množice smo tako ocenjevali na treh različnih poskusih. Pri vsakem izmed poskusov smo za simulacijo odvisnosti nastanka manjkajočih vrednosti izbrali drug neodvisni atribut, pri čemer smo opravili pet ponovitev primerjav za vsakega izmed sedmih različnih deležev manjkajočih vrednosti. Rezultati se nahajajo v spodnjih tabelah in grafih.

Tabela 6.2-15: Povprečne absolutne razlike varianc pred in po nadomeščanju manjkajočih vrednosti razrednega atributa umetne podatkovne množice, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 1. neodvisnega atributa

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	2,23E-4 ± 1,70E-4	4,74E-4 ± 2,38E-4	7,61E-4 ± 1,20E-4	1,39E-3 ± 7,16E-4	5,40E-4 ± 6,32E-4	3,29E-3 ± 6,01E-4	8,51E-3 ± 1,47E-3
Modelno drevo	1,35E-4 ± 1,26E-4	4,85E-4 ± 2,00E-4	8,67E-4 ± 4,25E-4	1,60E-3 ± 6,93E-4	3,19E-3 ± 3,93E-4	4,54E-3 ± 6,44E-6	9,59E-3 ± 5,53E-4
MNSK	1,36E-4 ± 6,83E-5	5,46E-4 ± 1,20E-4	1,12E-3 ± 1,47E-4	1,86E-3 ± 5,49E-5	2,42E-3 ± 5,88E-5	2,89E-3 ± 1,69E-4	6,61E-3 ± 9,00E-6
Lin. Reg.	2,52E-4 ± 7,37E-5	7,12E-4 ± 1,22E-4	1,25E-3 ± 9,82E-5	1,80E-3 ± 2,62E-5	2,45E-3 ± 3,84E-5	2,91E-3 ± 2,46E-5	6,58E-3 ± 8,84E-6
K-NN	1,36E-4 ± 1,23E-4	2,96E-4 ± 3,07E-4	2,52E-4 ± 2,02E-4	1,75E-3 ± 2,48E-4	3,09E-3 ± 1,05E-4	5,21E-3 ± 7,61E-5	1,04E-2 ± 8,92E-6
ZeroR	3,36E-4 ± 1,33E-6	1,64E-3 ± 6,38E-6	3,23E-3 ± 2,44E-5	4,57E-3 ± 2,46E-5	5,92E-3 ± 3,44E-5	7,19E-3 ± 5,11E-6	1,28E-2 ± 8,92E-7
Vstavi 0	1,55E-3 ± 1,62E-6	7,15E-3 ± 2,45E-5	1,29E-2 ± 3,37E-5	1,75E-2 ± 3,60E-5	2,08E-2 ± 3,22E-5	2,33E-2 ± 3,38E-5	2,06E-2 ± 4,35E-6
Bagging	1,18E-4 ± 1,07E-4	5,79E-4 ± 3,42E-4	7,84E-4 ± 5,29E-4	2,11E-3 ± 4,32E-4	3,05E-3 ± 4,09E-4	5,23E-3 ± 3,10E-4	1,12E-2 ± 8,92E-5
Rot. reg. gozd	1,33E-4 ± 7,57E-5	5,40E-4 ± 1,24E-4	1,05E-3 ± 1,37E-4	1,81E-3 ± 4,75E-5	2,42E-3 ± 6,19E-5	3,15E-3 ± 6,90E-5	6,72E-3 ± 1,57E-4
RRG var. 1	1,31E-4 ± 7,37E-5	5,24E-4 ± 1,32E-4	1,02E-3 ± 1,35E-4	1,74E-3 ± 4,58E-5	2,35E-3 ± 4,57E-5	2,98E-3 ± 6,38E-5	6,56E-3 ± 6,84E-5
RRG var. 2	4,10E-5 ± 7,07E-5	9,94E-5 ± 2,02E-4	4,85E-4 ± 2,13E-4	1,28E-3 ± 6,40E-5	1,89E-3 ± 5,89E-5	2,31E-3 ± 3,05E-5	5,54E-3 ± 1,73E-4
SPSS	1,10E-4 ± 2,26E-5	6,37E-4 ± 4,72E-4	1,33E-3 ± 2,88E-4	5,85E-4 ± 4,75E-4	1,89E-3 ± 3,60E-6	2,58E-3 ± 1,06E-3	1,76E-3 ± 3,23E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

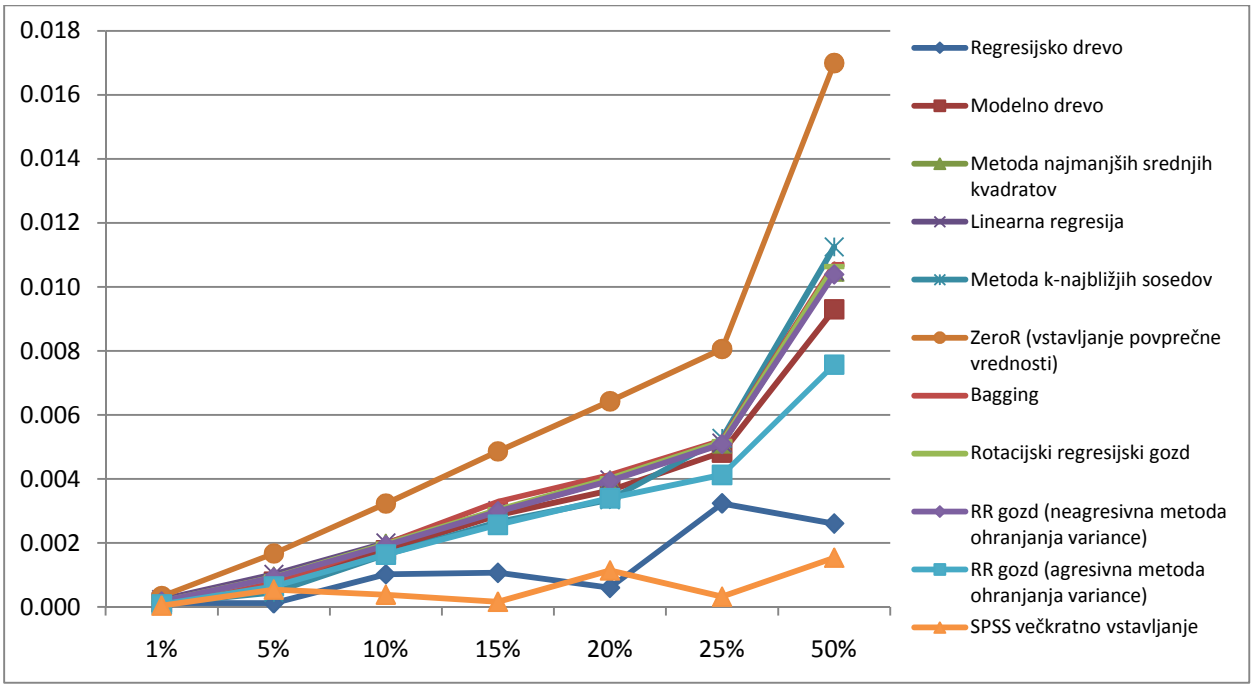


Graf 6.2-9: Povprečne absolutne razlike varianc vrednosti razrednega atributa pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 1. atributa (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene razlike varianc močno odstopajo)

Tabela 6.2-16: Povprečne absolutne razlike varianc pred in po nadomeščanju manjkajočih vrednosti razrednega atributa umetne podatkovne množice, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 2. neodvisnega atributa

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	1,20E-4 ± 1,29E-4	1,25E-4 ± 8,95E-5	1,02E-3 ± 4,30E-4	1,07E-3 ± 9,38E-4	6,03E-4 ± 3,52E-4	3,24E-3 ± 5,53E-4	2,61E-3 ± 1,29E-4
Modelno drevo	1,48E-4 ± 9,22E-5	7,93E-4 ± 1,03E-4	1,77E-3 ± 1,69E-4	2,88E-3 ± 5,88E-5	3,64E-3 ± 3,18E-4	4,83E-3 ± 1,02E-4	9,30E-3 ± 1,63E-4
MNSK	1,77E-4 ± 5,86E-5	9,50E-4 ± 9,27E-5	1,98E-3 ± 7,95E-5	2,99E-3 ± 2,37E-5	3,96E-3 ± 1,01E-4	5,11E-3 ± 3,57E-5	1,05E-2 ± 5,23E-5
Lin. Reg.	2,31E-4 ± 6,08E-5	1,03E-3 ± 8,96E-5	2,01E-3 ± 5,59E-5	3,01E-3 ± 1,96E-5	3,97E-3 ± 3,69E-5	5,12E-3 ± 2,11E-5	1,05E-2 ± 4,61E-5
K-NN	1,13E-4 ± 6,57E-5	4,56E-4 ± 2,41E-4	1,63E-3 ± 1,53E-4	2,65E-3 ± 2,46E-4	3,35E-3 ± 2,46E-4	5,27E-3 ± 1,31E-4	1,12E-2 ± 1,33E-4
ZeroR	3,37E-4 ± 2,49E-6	1,67E-3 ± 1,27E-5	3,23E-3 ± 1,09E-5	4,86E-3 ± 1,54E-5	6,43E-3 ± 2,88E-5	8,06E-3 ± 1,89E-5	1,70E-2 ± 1,91E-5
Vstavi 0	1,55E-3 ± 2,95E-6	7,28E-3 ± 4,32E-5	1,34E-2 ± 8,18E-5	1,86E-2 ± 4,90E-5	2,28E-2 ± 5,10E-5	2,62E-2 ± 4,45E-5	2,77E-2 ± 4,88E-5
Bagging	1,62E-4 ± 7,55E-5	7,32E-4 ± 2,73E-4	1,97E-3 ± 8,71E-5	3,29E-3 ± 2,90E-4	4,12E-3 ± 2,32E-4	5,20E-3 ± 1,45E-4	1,07E-2 ± 3,43E-4
Rot. reg. gozd	1,74E-4 ± 6,81E-5	9,35E-4 ± 1,01E-4	1,96E-3 ± 7,19E-5	3,04E-3 ± 2,16E-5	4,01E-3 ± 8,07E-5	5,16E-3 ± 7,83E-5	1,06E-2 ± 3,12E-4
RRG var. 1	1,72E-4 ± 6,68E-5	9,20E-4 ± 1,00E-4	1,92E-3 ± 6,39E-5	2,99E-3 ± 3,30E-5	3,94E-3 ± 6,57E-5	5,09E-3 ± 8,50E-5	1,04E-2 ± 2,44E-4
RRG var. 2	7,92E-5 ± 6,51E-5	6,42E-4 ± 8,74E-5	1,64E-3 ± 5,43E-5	2,57E-3 ± 1,11E-4	3,40E-3 ± 2,52E-5	4,13E-3 ± 1,49E-4	7,57E-3 ± 2,64E-4
SPSS	4,14E-5 ± 4,21E-5	5,45E-4 ± 5,67E-4	3,85E-4 ± 3,96E-4	1,62E-4 ± 4,18E-5	1,14E-3 ± 8,37E-4	3,21E-4 ± 3,07E-4	1,54E-3 ± 1,04E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

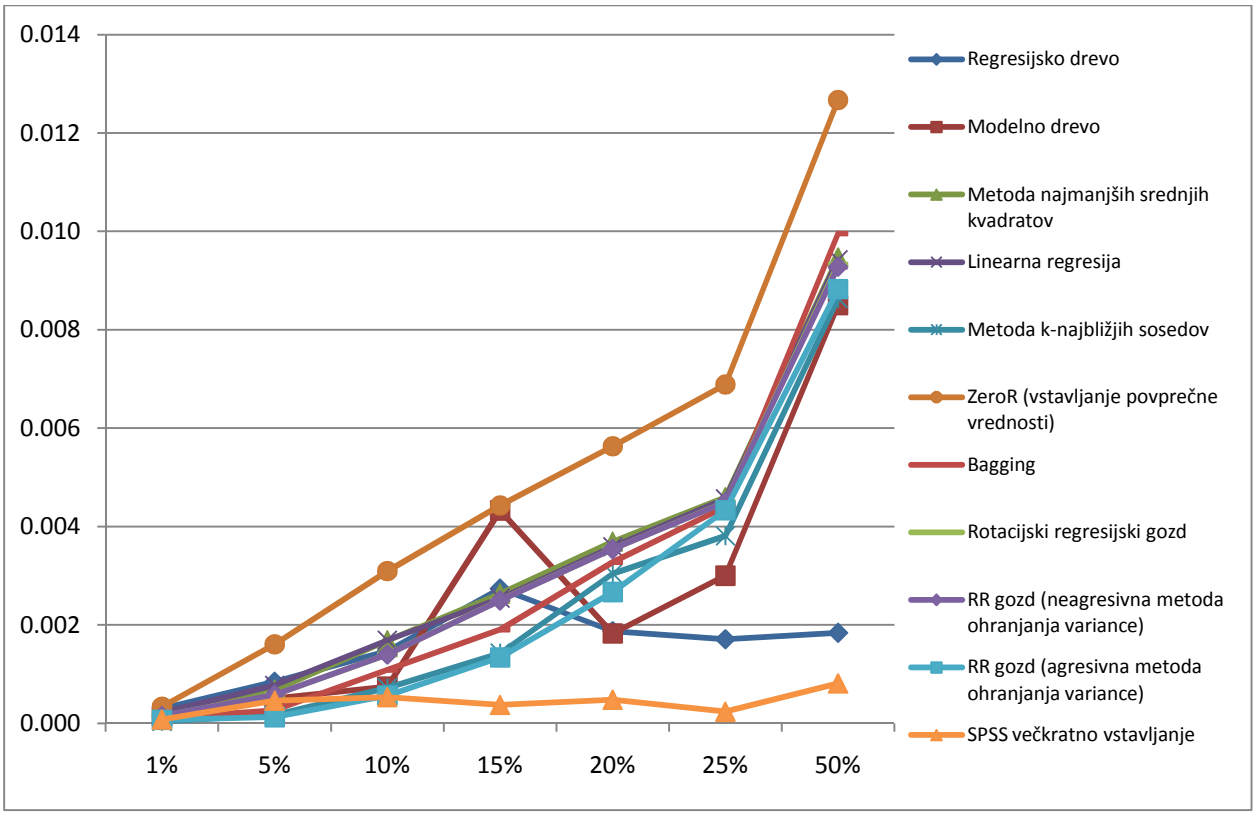


Graf 6.2-10: Povprečne absolutne razlike varianc vrednosti razrednega atributa pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 2. atributa (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene razlike varianc močno odstopajo)

Tabela 6.2-17: Povprečne absolutne razlike varianc pred in po nadomeščanju manjkajočih vrednosti razrednega atributa umetne podatkovne množice, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 3. neodvisnega atributa

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	2,95E-4 ± 2,46E-4	8,42E-4 ± 1,20E-3	1,47E-3 ± 1,00E-3	2,74E-3 ± 5,18E-4	1,87E-3 ± 1,07E-3	1,71E-3 ± 8,93E-4	1,84E-3 ± 1,35E-4
Modelno drevo	1,41E-4 ± 7,69E-5	5,04E-4 ± 3,02E-4	7,47E-4 ± 6,23E-4	4,33E-3 ± 7,56E-4	1,83E-3 ± 8,07E-4	3,00E-3 ± 4,10E-4	8,50E-3 ± 1,35E-5
MNSK	1,61E-4 ± 4,71E-5	6,76E-4 ± 5,25E-5	1,69E-3 ± 1,22E-4	2,64E-3 ± 9,65E-5	3,69E-3 ± 8,54E-5	4,60E-3 ± 6,98E-5	9,47E-3 ± 1,35E-6
Lin. Reg.	2,18E-4 ± 4,63E-5	7,71E-4 ± 6,40E-5	1,69E-3 ± 8,03E-5	2,54E-3 ± 2,37E-5	3,59E-3 ± 4,39E-5	4,56E-3 ± 2,66E-5	9,42E-3 ± 1,33E-6
K-NN	3,97E-5 ± 3,41E-5	1,67E-4 ± 1,04E-4	7,18E-4 ± 2,61E-4	1,43E-3 ± 1,74E-4	3,04E-3 ± 6,18E-5	3,81E-3 ± 8,80E-5	8,64E-3 ± 1,35E-6
ZeroR	3,37E-4 ± 9,45E-7	1,61E-3 ± 2,27E-5	3,09E-3 ± 1,86E-5	4,43E-3 ± 1,90E-5	5,63E-3 ± 8,00E-6	6,89E-3 ± 6,98E-6	1,27E-2 ± 1,33E-6
Vstavi 0	1,55E-3 ± 1,78E-6	7,10E-3 ± 2,19E-5	1,27E-2 ± 2,77E-5	1,71E-2 ± 2,79E-5	2,04E-2 ± 3,44E-5	2,27E-2 ± 3,66E-5	2,27E-2 ± 4,65E-7
Bagging	1,01E-4 ± 7,90E-5	2,62E-4 ± 3,13E-4	1,08E-3 ± 2,98E-4	1,90E-3 ± 5,82E-4	3,28E-3 ± 4,31E-4	4,40E-3 ± 3,31E-4	9,95E-3 ± 1,34E-5
Rot. reg. gozd	1,48E-4 ± 5,74E-5	5,76E-4 ± 9,53E-5	1,40E-3 ± 2,88E-4	2,50E-3 ± 2,58E-5	3,55E-3 ± 4,89E-5	4,50E-3 ± 2,72E-5	9,33E-3 ± 1,00E-5
RRG var. 1	1,48E-4 ± 5,74E-5	5,76E-4 ± 9,54E-5	1,40E-3 ± 2,88E-4	2,50E-3 ± 2,59E-5	3,54E-3 ± 5,04E-5	4,49E-3 ± 2,73E-5	9,28E-3 ± 4,03E-5
RRG var. 2	7,17E-5 ± 6,03E-5	1,26E-4 ± 1,57E-4	5,65E-4 ± 6,78E-4	1,34E-3 ± 6,46E-5	2,67E-3 ± 6,58E-5	4,33E-3 ± 3,91E-5	8,83E-3 ± 2,27E-4
SPSS	7,65E-5 ± 4,93E-5	4,61E-4 ± 1,96E-4	5,34E-4 ± 4,44E-4	3,74E-4 ± 3,76E-5	4,81E-4 ± 2,86E-4	2,36E-4 ± 1,13E-4	8,12E-4 ± 7,20E-5

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**



Graf 6.2-11: Povprečne absolutne razlike varianc vrednosti razrednega atributa pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je nastanek manjkajočih vrednosti odvisen od vrednosti 3. atributa (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene razlike varianc močno odstopajo)

Iz grafov lahko razberemo, da se stopnja ohranjanja variance opazno slabša z večanjem deleža manjkajočih vrednosti. Edina izjema je metoda programskega paketa SPSS, ki dobro ohranja varianco ne glede na stopnjo manjkajočih vrednosti in je občutno boljša od preostalih metod. Daleč najslabši metodi sta, pričakovano, vstavljanje ničel in vstavljanje povprečnih vrednosti, medtem ko se je osnovna različica rotacijskega regresijskega gozda obnesla podobno kot metoda najmanjših srednjih kvadratov, linearna regresija ter bagging. Rotacijski regresijski gozd z neagresivno metodo ohranjanja variance tudi sodi v to skupino in je le malenkost boljši od osnovne različice. Varianta z agresivno metodo se je odrezala precej bolje in pri majhnih odstotkih manjkajočih vrednosti prekosila tudi metodo orodja za statistično analizo SPSS. Pri visokih deležih manjkajočih vrednosti je presenetljivo dobre rezultate doseglo regresijsko drevo, ki se pri prejšnjih poskusih ni izkazalo.

Kot že vsakič poprej smo tudi tokrat metode rangirali s pomočjo Friedmanovega testa. Rezultati testa in dobljeni povprečni rangi se nahajajo v tabelah (Tabela 6.2-18 in Tabela 6.2-19).

Tabela 6.2-18: Friedmanov test za primerjavo metod po sposobnosti ohranjanja variance pri nadomeščanju manjkajočih vrednostih tipa MAR

N	χ^2	stopnje prostosti	p
21	157,125	11	0,000

Tabela 6.2-19: Friedmanova razvrstitev metod po rangih glede na uspešnost ohranjanja variance pri nadomeščanju manjkajočih vrednosti tipa MAR

	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Povprečni rang*	4,33	5,29	7,62	8,19	4,81	11,00	12,00	7,05	7,14	5,57	2,67	2,33

*) Povprečni rang po Friedmanovem testu (N=21, χ^2=157,125, sp=11, sig.=0,000)

Test je potrdil našo domnevo, da se metode po sposobnosti ohranjanja variance signifikantno razlikujejo med seboj. Kot smo že razbrali iz grafov, je najbolje rangirana metoda orodja za statistično analizo SPSS, ki ji sledi različica našega rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance, predvsem po zaslugi dobrih rezultatov pri majhnih deležih manjkajočih vrednosti. Na tretjem mestu se nahaja regresijsko drevo, ki je dobro ohranilo varianco pri višjih odstotkih manjkajočih vrednosti. Pričakovano se je tudi različica rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance uvrstila višje kot osnovni rotacijski regresijski gozd.

Za preverjanje ničelne hipoteze, da se dve primerjani metodi bistveno ne razlikujeta po svoji sposobnosti ohranjanja variance, smo tudi tokrat uporabili Wilcoxonov test predznačenih rangov. Domnevali smo, da se bo kot signifikantno najučinkovitejša izkazala metoda orodja za statistično analizo SPSS, zato smo jo paroma primerjali z drugimi metodami. Enako smo naredili z rotacijskim regresijskim gozdom z agresivno metodo ohranjanja variance. Posebej smo primerjali tudi preostali različici rotacijskega regresijskega gozda, z namenom potrditve hipoteze, da je različica z dodatno metodo za ohranjanje variance signifikantno uspešnejša od osnovne različice. Rezultati analize, opravljene z orodjem za statistično analizo SPSS, so predstavljeni v spodnji tabeli.

Tabela 6.2-20: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo metod po uspešnosti ohranjanja variance pri nadomeščanju manjkajočih vrednosti tipa MAR (N=21)

1. metoda	2. metoda	Z	Signifikanca
Metoda orodja za statistično analizo SPSS	Regresijsko drevo	-2,624[b]	0,009
Metoda orodja za statistično analizo SPSS	Modelno drevo	-3,563[b]	0,000
Metoda orodja za statistično analizo SPSS	Metoda najmanjših srednjih kvadratov	-3,736[b]	0,000
Metoda orodja za statistično analizo SPSS	Linearna regresija	-3,945[b]	0,000
Metoda orodja za statistično analizo SPSS	Metoda k-najbližjih sosedov	-3,041[b]	0,002
Metoda orodja za statistično analizo SPSS	Vstavljanje povprečne vrednosti	-4,015[b]	0,000
Metoda orodja za statistično analizo SPSS	Vstavljanje ničel	-4,015[b]	0,000
Metoda orodja za statistično analizo SPSS	Bagging	-3,458[b]	0,001
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd	-3,702[b]	0,000
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-3,702[b]	0,000
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-2,589[b]	0,010
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Regresijsko drevo	-0,921[a]	0,357
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Modelno drevo	-2,555[b]	0,011
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Linearna regresija	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda k-najbližjih sosedov	-1,999[b]	0,046
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Vstavljanje povprečne vrednosti	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Vstavljanje ničel	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Bagging	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-4,015[b]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-4,015[b]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-4,015[b]	0,000

[a] na podlagi negativnih rangov
[b] na podlagi pozitivnih rangov

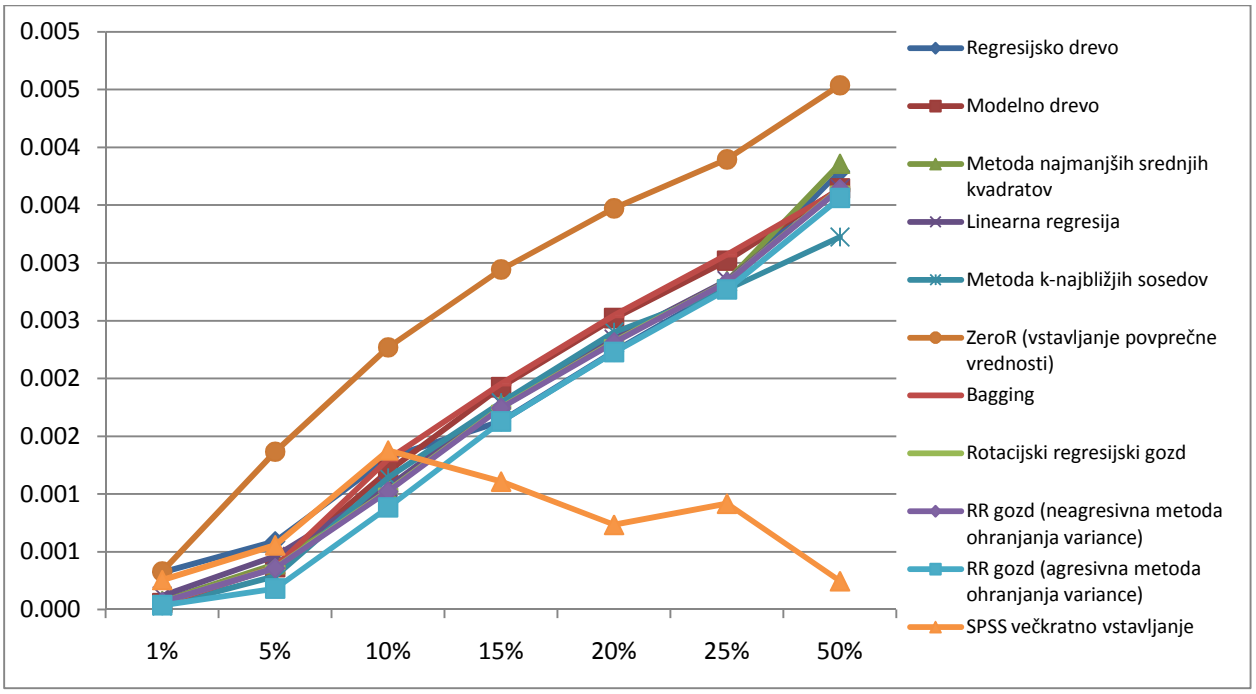
Ob upoštevanju pogoja $z < -1{,}96$ pri $p < 0{,}05$ je razvidno, da je metoda orodja za statistično analizo SPSS signifikantno boljša od drugih metod. Še najbližje ji je različica rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance, ki je signifikantno uspešnejša od vseh preostalih metod, z izjemo regresijskega drevesa. Dobri rezultati regresijskega drevesa so kar nekoliko presenetljivi, če upoštevamo, da je ta ista metoda bila med slabšimi pri ovrednotenju uspešnosti ohranjanja variance na poskusih z manjkajočimi vrednostmi tipa MCAR. Klub majhnim razlikam pri ohranjanju variance na posameznih poskusih lahko trdimo tudi, da se je različica rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance izkazala kot statistično signifikantno uspešnejša od osnovne variante regresijskega gozda.

Pri ovrednotenju sposobnosti ohranjanja variance ob nadomeščanju manjkajočih vrednosti tipa NMAR smo, tako kot pri ocenjevanju natančnosti metod, opravili poskuse na umetni podatkovni množici in na množici *Concrete*, pri kateri smo za nadomeščanje izbrali vrednosti 8. atributa. Za vsako izmed sedmih stopenj manjkajočih vrednosti smo opravili po pet meritev. Povprečne ocene ohranjanja variance na obeh množicah se nahajajo v spodnjih tabelah in grafih.

Tabela 6.2-21: Povprečne absolutne razlike varianc pred in po nadomeščanju manjkajočih vrednosti razrednega atributa umetne podatkovne množice, kadar je mehanizem nastanka manjkajočih vrednosti NMAR

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	3,23E-4 ± 1,20E-4	5,91E-4 ± 2,16E-4	1,31E-3 ± 4,48E-4	1,63E-3 ± 2,82E-4	2,23E-3 ± 1,62E-4	2,80E-3 ± 9,28E-5	3,79E-3 ± 1,89E-4
Modelno drevo	5,59E-5 ± 5,12E-5	3,67E-4 ± 9,07E-5	1,20E-3 ± 9,46E-5	1,92E-3 ± 9,40E-5	2,52E-3 ± 4,61E-5	3,02E-3 ± 1,14E-5	3,64E-3 ± 4,25E-5
MNSK	7,31E-5 ± 2,87E-5	3,92E-4 ± 4,39E-5	1,03E-3 ± 3,54E-5	1,78E-3 ± 2,00E-5	2,33E-3 ± 3,57E-5	2,84E-3 ± 3,46E-6	3,86E-3 ± 3,81E-6
Lin. Reg.	1,16E-4 ± 2,75E-5	4,60E-4 ± 4,89E-5	1,06E-3 ± 3,29E-5	1,78E-3 ± 1,89E-5	2,34E-3 ± 1,22E-5	2,85E-3 ± 3,46E-6	3,64E-3 ± 4,09E-6
K-NN	3,09E-5 ± 2,22E-5	2,88E-4 ± 8,10E-5	1,14E-3 ± 1,21E-5	1,79E-3 ± 6,03E-5	2,40E-3 ± 2,23E-5	2,77E-3 ± 8,67E-6	3,22E-3 ± 2,93E-6
ZeroR	3,29E-4 ± 3,18E-6	1,37E-3 ± 1,75E-5	2,27E-3 ± 1,15E-5	2,94E-3 ± 1,27E-5	3,47E-3 ± 9,88E-6	3,90E-3 ± 1,52E-6	4,54E-3 ± 3,81E-6
Vstavi 0	1,54E-3 ± 7,11E-7	6,92E-3 ± 3,09E-6	1,21E-2 ± 3,46E-6	1,59E-2 ± 7,10E-6	1,85E-2 ± 8,50E-6	2,00E-2 ± 6,58E-7	1,64E-2 ± 5,52E-7
Bagging	4,91E-5 ± 3,37E-5	3,69E-4 ± 4,61E-5	1,30E-3 ± 1,09E-4	1,95E-3 ± 1,12E-4	2,54E-3 ± 7,80E-5	3,07E-3 ± 3,55E-5	3,63E-3 ± 3,21E-6
Rot. reg. gozd	6,50E-5 ± 2,99E-5	3,64E-4 ± 3,87E-5	1,03E-3 ± 3,24E-5	1,75E-3 ± 1,90E-5	2,32E-3 ± 1,74E-5	2,84E-3 ± 3,67E-6	3,65E-3 ± 3,28E-6
RRG var. 1	6,50E-5 ± 2,98E-5	3,58E-4 ± 3,92E-5	1,02E-3 ± 3,19E-5	1,74E-3 ± 1,92E-5	2,31E-3 ± 1,38E-5	2,83E-3 ± 3,76E-6	3,64E-3 ± 3,23E-6
RRG var. 2	3,77E-5 ± 2,46E-5	1,82E-4 ± 6,47E-5	8,83E-4 ± 4,98E-5	1,63E-3 ± 3,88E-5	2,23E-3 ± 2,15E-5	2,77E-3 ± 1,04E-5	3,56E-3 ± 1,90E-5
SPSS	2,56E-4 ± 1,47E-4	5,57E-4 ± 4,08E-4	1,38E-3 ± 4,71E-5	1,11E-3 ± 4,14E-5	7,34E-4 ± 2,92E-4	9,17E-4 ± 7,91E-4	2,46E-4 ± 1,14E-4

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**

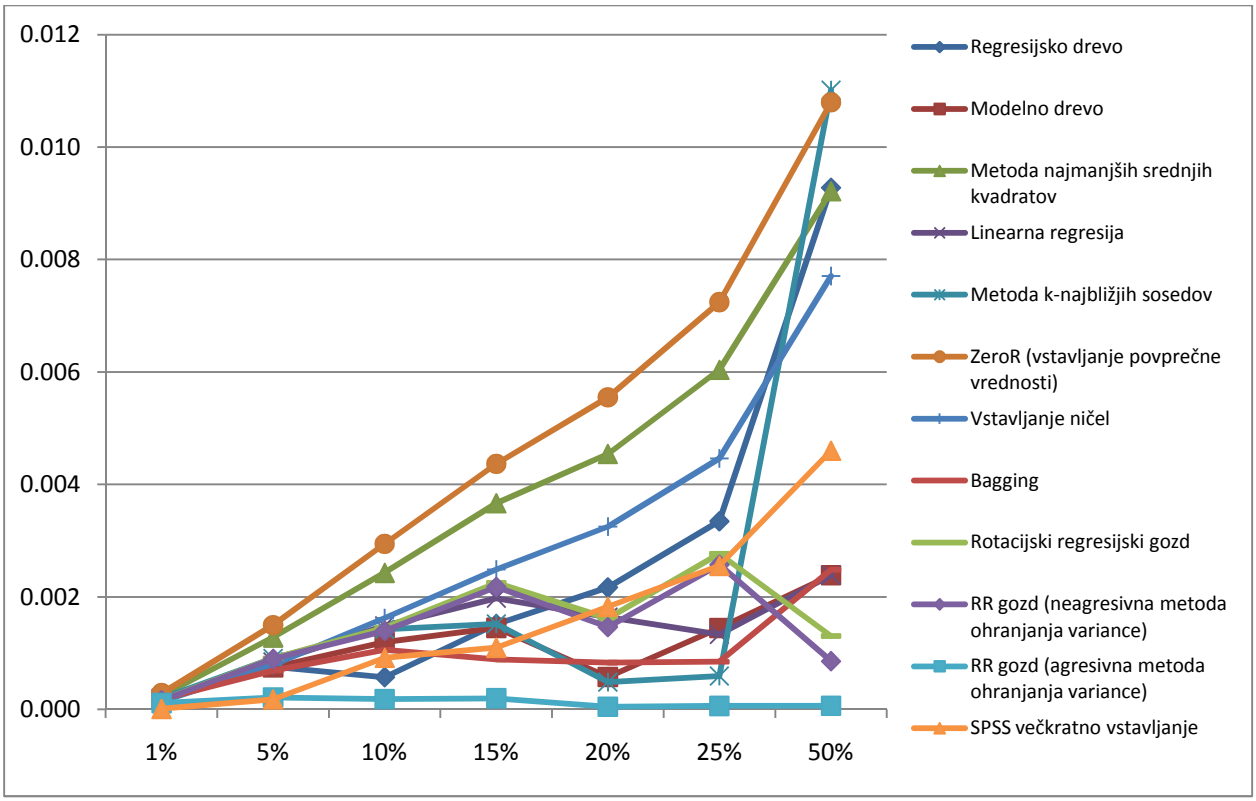


Graf 6.2-12: Povprečne absolutne razlike varianc vrednosti razrednega atributa pri različnih stopnjah manjkajočih vrednosti v umetni podatkovni množici, kadar je mehanizem nastanka manjkajočih vrednosti NMAR (metoda vstavljanja ničel zaradi preglednosti grafa ni vključena, saj njene povprečne napake močno odstopajo)

Tabela 6.2-22: Povprečne absolutne razlike varianc pred in po nadomeščanju manjkajočih vrednosti 8. atributa podatkovne množice *Concrete*, kadar je mehanizem nastanka manjkajočih vrednosti NMAR

Metoda*	1% MV	5% MV	10% MV	15% MV	20% MV	25% MV	50% MV
Reg. drevo	1,91E-4 ± 7,88E-5	7,57E-4 ± 1,70E-4	5,73E-4 ± 3,19E-4	1,52E-3 ± 9,71E-4	2,17E-3 ± 1,96E-3	3,35E-3 ± 1,38E-3	9,28E-3 ± 3,35E-3
Modelno drevo	1,92E-4 ± 7,96E-5	7,48E-4 ± 2,87E-4	1,20E-3 ± 2,54E-4	1,45E-3 ± 1,54E-4	5,77E-4 ± 5,05E-5	1,45E-3 ± 7,12E-4	2,39E-3 ± 1,22E-6
MNSK	2,43E-4 ± 7,45E-6	1,28E-3 ± 3,30E-5	2,43E-3 ± 3,13E-5	3,67E-3 ± 6,11E-5	4,55E-3 ± 3,02E-5	6,04E-3 ± 7,59E-5	9,21E-3 ± 8,17E-7
Lin. Reg.	1,84E-4 ± 4,02E-5	8,45E-4 ± 8,13E-5	1,48E-3 ± 8,37E-5	1,98E-3 ± 1,60E-4	1,64E-3 ± 1,41E-4	1,33E-3 ± 7,12E-5	2,39E-3 ± 4,58E-5
K-NN	1,98E-4 ± 6,15E-5	9,06E-4 ± 3,45E-4	1,42E-3 ± 4,14E-4	1,52E-3 ± 5,46E-4	4,84E-4 ± 1,22E-5	5,94E-4 ± 2,39E-4	1,10E-2 ± 3,65E-3
ZeroR	2,89E-4 ± 4,43E-6	1,50E-3 ± 3,88E-5	2,94E-3 ± 2,40E-5	4,37E-3 ± 8,55E-5	5,55E-3 ± 9,08E-7	7,24E-3 ± 1,06E-4	1,08E-2 ± 8,09E-4
Vstavi 0	1,45E-4 ± 2,87E-6	7,92E-4 ± 2,65E-5	1,62E-3 ± 1,31E-5	2,49E-3 ± 6,51E-5	3,25E-3 ± 2,09E-6	4,46E-3 ± 8,26E-5	7,71E-3 ± 0,00E+0
Bagging	1,82E-4 ± 9,13E-5	6,90E-4 ± 4,01E-4	1,06E-3 ± 9,12E-4	8,91E-4 ± 5,62E-4	8,35E-4 ± 9,26E-4	8,47E-4 ± 8,14E-4	2,48E-3 ± 3,32E-5
Rot. reg. gozd	1,69E-4 ± 8,27E-5	9,10E-4 ± 1,40E-4	1,45E-3 ± 3,79E-4	2,25E-3 ± 6,31E-4	1,62E-3 ± 2,76E-4	2,77E-3 ± 3,03E-4	1,30E-3 ± 9,02E-4
RRG var. 1	1,64E-4 ± 8,14E-5	8,99E-4 ± 1,42E-4	1,40E-3 ± 3,43E-4	2,18E-3 ± 6,16E-4	1,47E-3 ± 2,64E-4	2,57E-3 ± 3,06E-4	8,59E-4 ± 7,57E-4
RRG var. 2	1,14E-4 ± 6,28E-5	2,12E-4 ± 2,76E-4	1,86E-4 ± 2,30E-4	1,98E-4 ± 1,43E-4	4,66E-5 ± 3,20E-5	6,23E-5 ± 2,54E-5	6,46E-5 ± 6,81E-5
SPSS	1,16E-5 ± 1,22E-5	1,81E-4 ± 1,01E-4	9,19E-4 ± 3,35E-4	1,10E-3 ± 1,12E-3	1,83E-3 ± 8,23E-4	2,56E-3 ± 5,60E-4	4,60E-3 ± 1,44E-3

*) **Reg. drevo: Regresijsko drevo; MNSK: Metoda najmanjših srednjih kvadratov; Lin. Reg.: Linearna regresija; K-NN: Metoda k-najbližjih sosedov; ZeroR: Vstavljanje povprečne vrednosti; Rot. reg. gozd: Rotacijski regresijski gozd; RRG var. 1: Rotacijski regresijski gozd z neagresivno metodo za izboljšanje ohranjanja variance; RRG var. 2: Rotacijski regresijski gozd z agresivno metodo za izboljšanje ohranjanja variance**



Graf 6.2-13: Povprečne absolutne razlike varianc vrednosti 8. atributa pri različnih stopnjah manjkajočih vrednosti v podatkovni množici *Concrete*, kadar je mehanizem nastanka manjkajočih vrednosti NMAR

Že bežen pogled na oba grafa nam razkrije, da so razlike med metodami prišle precej bolj do izraza pri nadomeščanju manjkajočih vrednosti 8. atributa množice *Concrete*. V primeru umetne podatkovne množice je sposobnost ohranjanja variance pri vseh metodah, z izjemo metode orodja SPSS, občutno padala z večanjem deleža manjkajočih vrednosti, medtem ko je pri podatkovni množici *Concrete* več kot polovica vseh metod relativno dobro ohranila varianco vrednosti 8. atributa tudi pri 25% manjkajočih vrednosti. Razlog za to lahko spet poiščemo v močno asimetrični porazdelitvi vrednosti 8. atributa te množice. V takšnih pogojih se je najbolje znašla varianta rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance, ki je po nadomeščanju manjkajočih vrednosti obdržala varianco skoraj nespremenjeno, ne glede na delež manjkajočih vrednosti. Na umetni podatkovni množici je dobre rezultate dosegla le metoda statističnega orodja SPSS, izmed preostalih, »slabih« metod sta se najbolje odrezali obe različici rotacijskega regresijskega gozda z metodo za ohranjanje variance.

Zaradi majhnega števila primerjav na podlagi povprečnih ocen (7) smo pri rangiranju metod s pomočjo Friedmanovega testa uporabili ocene napak vseh 5 ponovitev poskusov in tako pri vsaki

podatkovni množici dobili 35 neodvisnih primerjav. Povprečni rangi in rezultati Friedmanovih testov se nahajajo v tabelah (Tabela 6.2-23 in Tabela 6.2-24).

Tabela 6.2-23: Friedmanova razvrstitev metod po rangih glede sposobnost ohranjanja variance pri napovedovanju manjkajočih vrednosti tipa NMAR v umetni podatkovni množici

	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Povprečni rang*	6,54	7,76	7,03	6,36	4,40	10,97	12,00	7,20	5,36	4,07	1,94	4,37

*) **Povprečni rang po Friedmanovem testu (N=35, χ^2=241,713, sp=11, sig.=0,000)**

Tabela 6.2-24: Friedmanov test za primerjavo metod po sposobnosti ohranjanja variance pri napovedovanju manjkajočih vrednosti tipa NMAR v umetni podatkovni množici

N	χ^2	stopnje prostosti	p
35	241,713	11	0,000

Tabela 6.2-25: Friedmanova razvrstitev metod po rangih glede sposobnost ohranjanja variance pri napovedovanju manjkajočih vrednosti tipa NMAR v podatkovni množici *Concrete*

	RD	MD	LMS	LR	KNN	Pov.	0	Bag.	RRG	RRG1	RRG2	SPSS
Povprečni rang*	6,40	5,04	10,54	6,13	6,37	11,80	8,06	4,91	6,97	5,83	1,43	4,51

*) **Povprečni rang po Friedmanovem testu (N=35, χ^2=220,960, sp=11, sig.=0,000)**

Tabela 6.2-26: Friedmanov test za primerjavo metod po sposobnosti ohranjanja variance pri napovedovanju manjkajočih vrednosti tipa NMAR v podatkovni množici *Concrete*

N	χ^2	stopnje prostosti	p
35	220,960	11	0,000

Če bi sodili na podlagi Friedmanovih rangov, bi lahko sklepali, da na umetni podatkovni množici varianco najbolje ohranjata različici rotacijskega regresijskega gozda z metodama za ohranjanje variance. Graf 6.2-12 sugerira drugače, razlog za slabši rang metode orodja za statistično analizo SPSS pa leži v slabi uvrstitvi metode pri majhnih odstotkih manjkajočih vrednosti. S pomočjo Wilcoxonovega testa predznačenih rangov smo primerjali obe najbolje rangirani metodi s preostalimi ter vse tri različice rotacijskega regresijskega gozda med seboj (Tabela 6.2-27).

Tabela 6.2-27: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo metod po uspešnosti ohranjanja variance pri nadomeščanju manjkajočih vrednosti tipa NMAR v umetni podatkovni množici (N=35)

1. metoda	2. metoda	Z	Signifikanca
Metoda orodja za statistično analizo SPSS	Regresijsko drevo	-4,375[a]	0,000
Metoda orodja za statistično analizo SPSS	Modelno drevo	-3,276[a]	0,001
Metoda orodja za statistično analizo SPSS	Metoda najmanjših srednjih kvadratov	-3,243[a]	0,001
Metoda orodja za statistično analizo SPSS	Linearna regresija	-3,243[a]	0,001
Metoda orodja za statistično analizo SPSS	Metoda k-najbližjih sosedov	-3,162[a]	0,002
Metoda orodja za statistično analizo SPSS	Vstavljanje povprečne vrednosti	-5,143[a]	0,000
Metoda orodja za statistično analizo SPSS	Vstavljanje ničel	-5,160[a]	0,000
Metoda orodja za statistično analizo SPSS	Bagging	-3,407[a]	0,001
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd	-3,227[a]	0,001
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-3,227[a]	0,001
Metoda orodja za statistično analizo SPSS	Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	-2,999[a]	0,003
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Regresijsko drevo	-5,168[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Modelno drevo	-5,078[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-5,161[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Linearna regresija	-5,160[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda k-najbližjih sosedov	-2,032[a]	0,042
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Vstavljanje povprečne vrednosti	-5,161[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Vstavljanje ničel	-5,161[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Bagging	-5,078[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-5,086[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-5,086[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-4,860[b]	0,000

[a] na podlagi pozitivnih rangov
[b] na podlagi negativnih rangov

Če upoštevamo pogoj $z < -1{,}96$ pri $p < 0{,}05$, ugotovimo, da je pri ohranjanju variance ob nadomeščanju manjkajočih vrednosti tipa NMAR v umetni podatkovni množici metoda orodja za statistično analizo SPSS signifikantno uspešnejša od drugih metod. Podobno je različica rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance signifikantno boljša od vseh preostalih metod. Kakor že vsakič dozdaj, lahko tudi tokrat trdimo, da je varianta rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance signifikantno uspešnejša od različice brez dodatne metode za ohranjanje variance.

Pri poskusih, opravljenih na podatkovni množici *Concrete*, izstopa rotacijski regresijski gozd z agresivno metodo ohranjanja variance. V primerjavo s pomočjo Wilcoxonovega testa smo vključili metodo orodja SPSS, ki je po Friedmanovem rangu na 2. mestu ter preostali različici rotacijskega regresijskega gozda (Tabela 6.2-28).

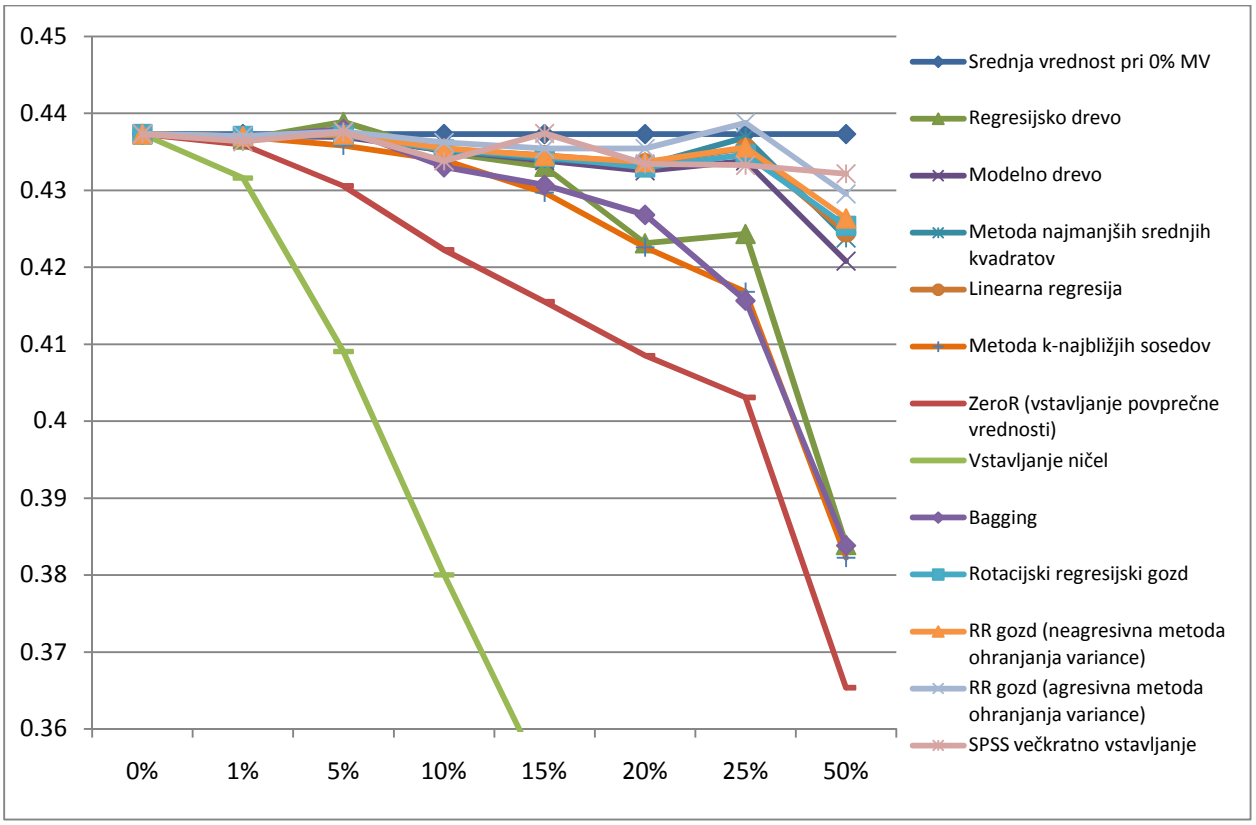
Tabela 6.2-28: Rezultati Wilcoxonovega testa predznačenih rangov za primerjavo metod po uspešnosti ohranjanja variance pri nadomeščanju manjkajočih vrednosti tipa NMAR v podatkovni množici *Concrete* (N=35)

1. metoda	2. metoda	Z	Signifikanca
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Regresijsko drevo	-5,061[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Modelno drevo	-5,159[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda najmanjših srednjih kvadratov	-5,159[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Linearna regresija	-5,159[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda k-najbližjih sosedov	-5,159[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Vstavljanje povprečne vrednosti	-5,159[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Vstavljanje ničel	-5,094[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Bagging	-5,045[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-5,086[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	-5,086[a]	0,000
Rotacijski regresijski gozd (agresivna metoda za ohranjanje variance)	Metoda orodja za statistično analizo SPSS	-4,324[a]	0,000
Rotacijski regresijski gozd (neagresivna metoda za ohranjanje variance)	Rotacijski regresijski gozd	-5,086[a]	0,000
Metoda orodja za statistično analizo SPSS	Bagging	-2,179[b]	0,029
Metoda orodja za statistično analizo SPSS	Modelno drevo	-1,474[b]	0,140

[a] na podlagi pozitivnih rangov
[b] na podlagi negativnih rangov

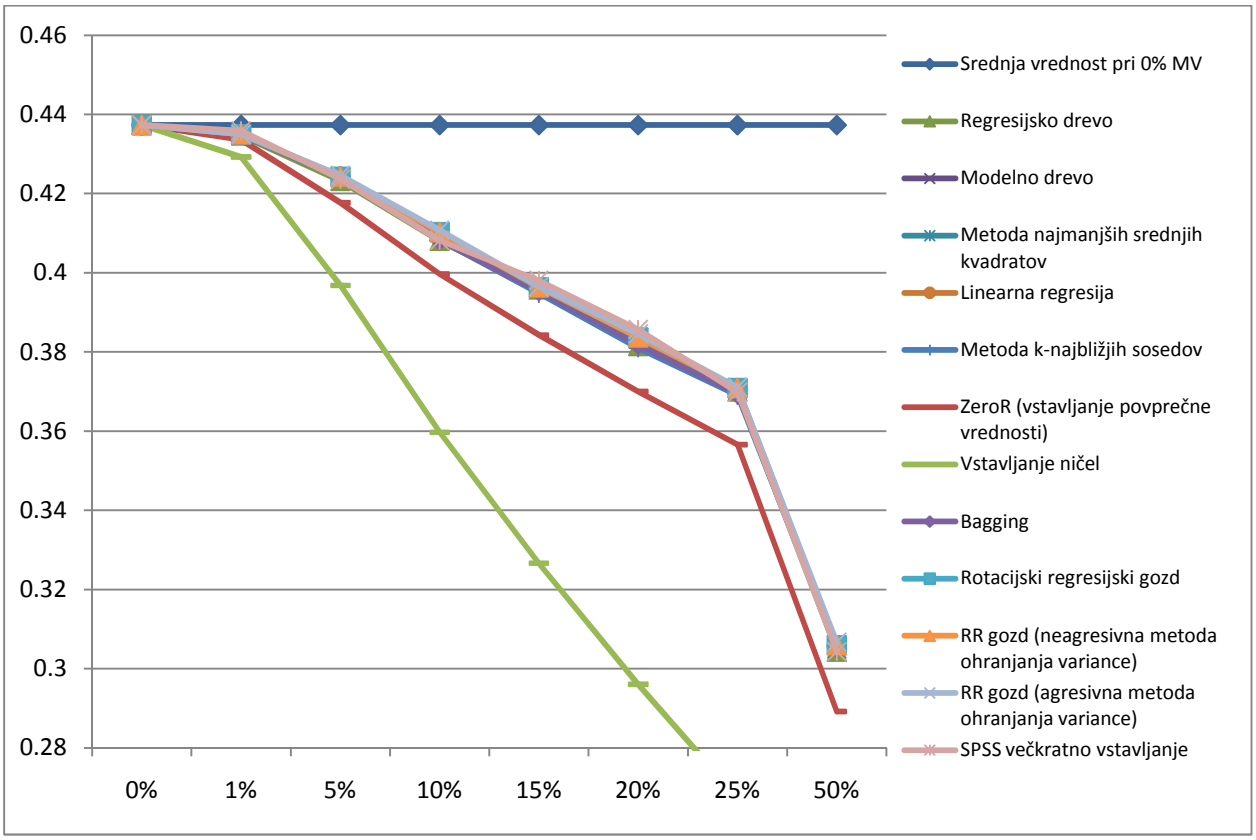
Pri ohranjanju variance ob nadomeščanju manjkajočih vrednosti v podatkovni množici *Concrete* se je rotacijski regresijski gozd z agresivno metodo ohranjanja variance pričakovano izkazal za statistično signifikantno najuspešnejšo metodo. Tretja najbolje rangirana metoda, bagging, je bila na podlagi Wilcoxonovega testa signifikantno boljša od metode SPSS, ki se je po Friedmanovem rangu uvrstila na 2. mesto. Rotacijski regresijski gozd z neagresivno metodo za ohranjanje variance se je še enkrat izkazal kot uspešnejši od variante brez te dodatne metode.

Posamična metoda se pri napovedovanju manjkajočih vrednosti lahko zanaša samo na preostale vrednosti v podatkovni množici. Zaradi tega sta srednja vrednost in pripadajoči standardni odklon vseh vrednosti določenega atributa po nadomeščanju manjkajočih vrednosti odvisna od srednje vrednosti in standardnega odklona pred nadomeščanjem. Kot smo to naredili že pri mehanizmu MCAR, si tudi za manjkajoče vrednosti tipa MAR in NMAR oglejmo, kako se spreminja srednja vrednost izbranega atributa s spreminjanjem deleža manjkajočih vrednosti. Zaradi čim bolj nazorne predstavitve vpliva mehanizma nastanka manjkajočih vrednosti smo kot primer izbrali umetno podatkovno množico, pri kateri so vrednosti razrednega atributa približno normalno porazdeljene (Graf 6.2-14 in Graf 6.2-15).



Graf 6.2-14: Vpliv deleža manjkajočih vrednosti tipa MAR na srednjo vrednost razrednega atributa umetne podatkovne množice po nadomeščanju z različnimi metodami

Če upoštevamo, da je srednja vrednost razrednega atributa pred nadomeščanjem manjkajočih vrednosti enaka srednji vrednosti po nadomeščanju z metodo vstavljanja povprečne vrednosti, lahko na grafu opazimo, da se srednja vrednost atributa skoraj linearno zmanjšuje z naraščanjem odstotka manjkajočih vrednosti. Kljub temu večina metod nima težav pri ohranjanju začetne srednje vrednosti (0,437). Vse metode, z izjemo vstavljanja ničel in vstavljanja povprečne vrednosti, še pri 10% manjkajočih vrednosti ohranijo srednjo vrednost znotraj odstopanja 1% (0,437 ± 0,004). Večina metod tudi pri 25% ohrani srednjo vrednost znotraj tega odstopanja, medtem ko pri 50% manjkajočih vrednosti to odstopanje naraste na še vedno sprejemljive 4% (0,437 ± 0,017), pri čemer se najbolje izkažejo metoda statističnega orodja SPSS (0,432) in vse tri različice rotacijskega regresijskega gozda (0,430; 0,426; 0,425). Regresijsko drevo, bagging in metoda k-najbližjih sosedov pri deležih manjkajočih vrednosti nad 10% odpovejo.

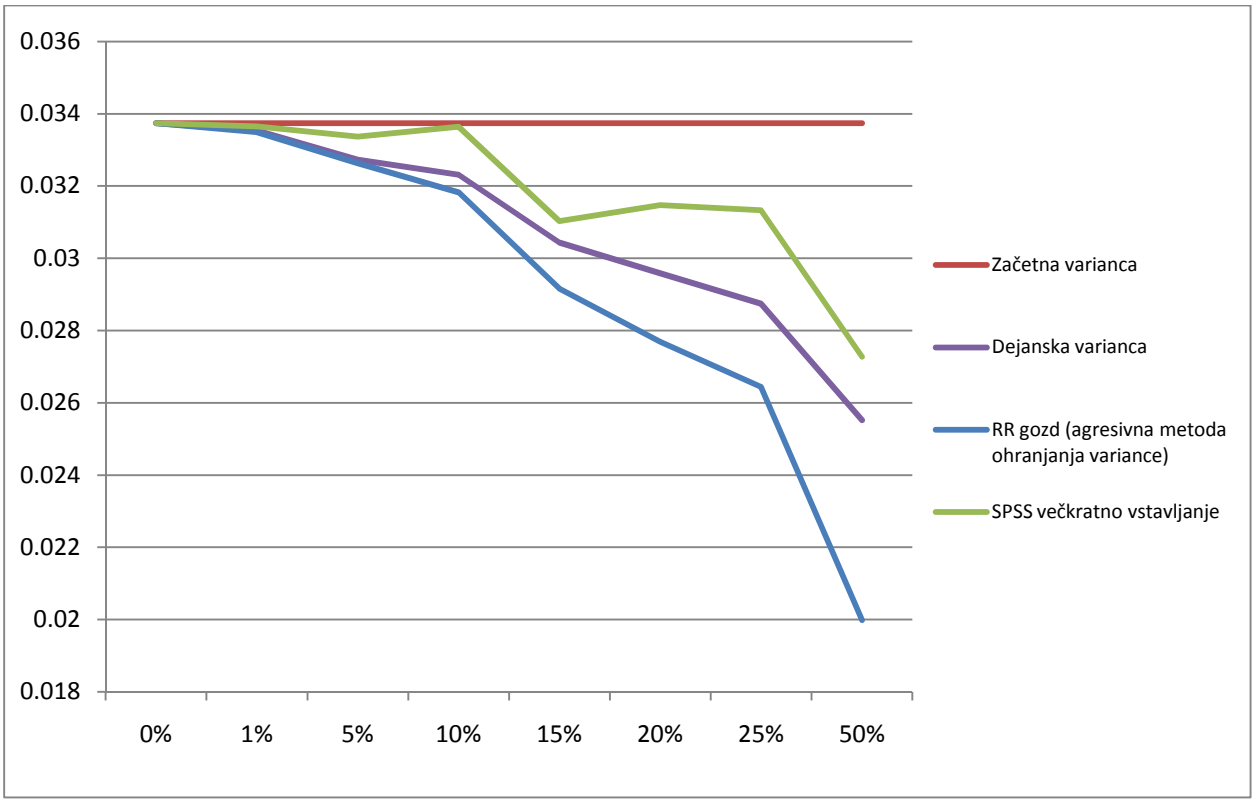


Graf 6.2-15: Vpliv deleža manjkajočih vrednosti tipa NMAR na srednjo vrednost razrednega atributa umetne podatkovne množice po nadomeščanju z različnimi metodami

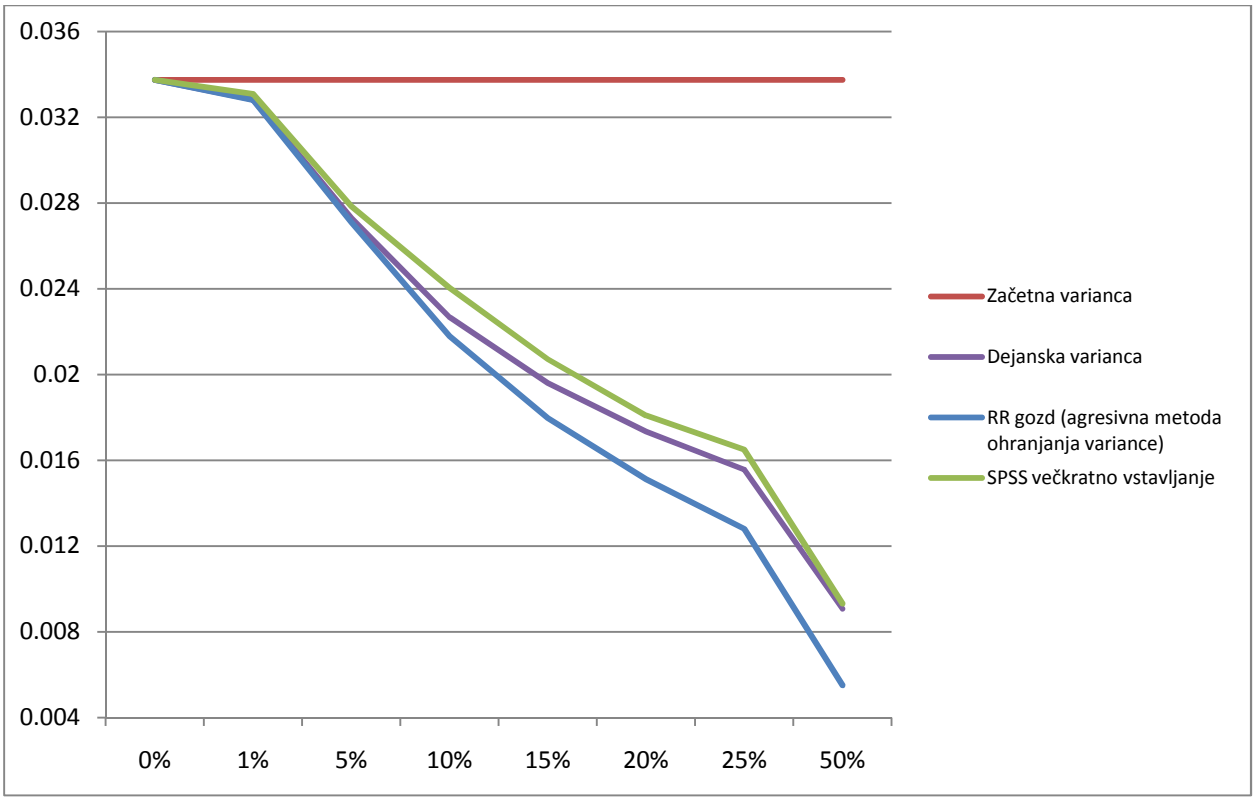
Situacija je povsem drugačna pri mehanizmu NMAR. Dejanska srednja vrednost pred nadomeščanjem strmo pada z naraščanjem manjkajočih vrednosti, po nadomeščanju pa med srednjimi vrednostmi skorajda ni razlik. Pri 1% manjkajočih vrednosti je povprečna srednja vrednost vseh metod

(brez vstavljanja povprečne vrednosti) 0,434 ± 0,002, torej znotraj 1% odstopanja, pri 15% manjkajočih vrednosti je 0,390 ± 0,021, kar je že več kot 10% odstopanje, medtem ko je pri 50% manjkajočih vrednosti povprečna srednja vrednost enaka 0,291 ± 0,049, torej za več kot tretjino manjša od začetne.

Sklepamo lahko, da imajo vse metode, kljub majhnim razlikam med njihovimi natančnostmi, velike težave pri napovedovanju manjkajočih vrednosti tipa NMAR. Podatkovna množica *Concrete*, ki smo jo uporabili za primerjavo z rezultati, pridobljenimi na umetno množici, je ena izmed izjem, ki potrjujejo to pravilo. To je tudi razlog, da se metode, ki temeljijo na večkratnem vstavljanju manjkajočih vrednosti, raje posvečajo ohranjanju srednje vrednosti in variance, kot pa sami natančnosti. Primer takšne metode je tudi metoda orodja za statistično analizo SPSS. Na predstavljenem primeru se ji sicer ni uspelo približati začetni srednji vrednosti, vendar je izmed uporabljenih metod edina, ki pri ohranjanju variance skrbi za ohranjanje ocene negotovosti. Medtem ko se pri vseh preostalih metodah z naraščanjem deleža manjkajočih vrednosti tipa MAR in NMAR praviloma zmanjšuje varianca (saj se zmanjšuje varianca preostalih, nemanjkajočih vrednosti), poskuša metoda SPSS varianco obdržati na dovolj visoki stopnji, ki ponazarja negotovost napovedi nadomestnih vrednosti (Graf 6.2-16 in Graf 6.2-17).



Graf 6.2-16: Varianca razrednega atributa umetne podatkovne množice pred in po nadomeščanju različnih deležev manjkajočih vrednosti tipa MAR (zaradi preglednosti sta prikazani samo dve najboljši metodi)



Graf 6.2-17: Varianca razrednega atributa umetne podatkovne množice pred in po nadomeščanju različnih deležev manjkajočih vrednosti tipa NMAR (zaradi preglednosti sta prikazani samo dve najboljši metodi)

Na primeru mehanizma MAR opazimo, da metoda orodja SPSS do vključno 10% deleža manjkajočih vrednosti zelo dobro ohranja začetno varianco razrednega atributa, medtem ko rotacijski regresijski gozd z agresivno metodo ohranjanja variance precej bolj natančno ohranja dejansko varianco atributa. Pri višjih odstotkih manjkajočih vrednosti se razlika med metodama poveča in obe se oddaljita tako od dejanske kot od začetne variance. Pri mehanizmu NMAR je razlika med metodama manj opazna, tudi tokrat pa je varianca razrednega atributa po nadomeščanju z metodo orodja za statistično analizo SPSS višja od dejanske.

6.3 Praktičen primer uporabe na podatkovni množici »Sindrom zapestnega prehoda«

Učinkovitost nadomeščanja manjkajočih vrednosti s pomočjo rotacijskega regresijskega gozda smo dodatno preverili tudi na praktičnem primeru. Pri eni izmed predhodnih raziskav smo namreč kreirali podatkovno bazo, sestavljeno iz numeričnih vrednosti, pridobljenih iz termografskih slik. Te numerične vrednosti predstavljajo temperature posameznih hrbtnih segmentov prostovoljcev in pacientov, pri katerih smo ugotovili utesnitev medianega živca v zapestnem prehodu fotografirane roke. V okviru raziskave smo izoblikovali diagnostično orodje, ki temelji na nevronskih mrežah in je sposobno z relativno visoko gotovostjo diagnosticirati hudo utesnitev živca - stanje v medicini imenovano »sindrom zapestnega prehoda« (tudi sindrom karpalnega kanala).

Tabela 6.3-1: Manjkajoče vrednosti v 8 nepopolnih vzorcih podatkovne množice »Sidrom zapestnega prehoda«

Segment	1	2	3	4	5	6	7	8
Zapestje pod palcem	✓	✓	✓	✓	✓	✓	✓	✓
Zapestje pod mezincem	✓	✓	✓	✓	✓	✓	✓	✓
Metakarpalni segment pod palcem	✓	✓	✓	✓	✓	✓	✓	✓
Metakarpalni segment pod kazalcem	✓	✓	✓	✓	✓	✓	✓	✓
Metakarpalni segment pod sredincem	✓	✓	✓	✓	✓	✓	✓	✓
Metakarpalni segment pod prstancem	✓	✓	✓	✓	✓	✓	✓	✓
Metakarpalni segment pod mezincem	✓	✓	✓	✓	✓	✓	✓	✓
Palec	✓	✓	✓	-	✓	✓	✓	✓
Kazalec	✓	✓	-	✓	✓	✓	-	✓
Sredinec	-	✓	-	-	✓	-	✓	✓
Prstanec	✓	-	✓	-	✓	✓	✓	-
Mezinec	✓	-	✓	✓	✓	✓	✓	-

V podatkovno bazo smo takrat vključili temperaturne meritve iz 251 termografskih slik, pri čemer se vsak vzorec množice sestoji iz 12 temperaturnih vrednosti, ki predstavljajo temperature različnih segmentov roke. Zaradi napak med fotografiranjem smo bili primorani izločiti večje število slik, pri katerih ni bilo možno opraviti segmentacije in posledično izračuna temperaturnih vrednosti posameznih segmentov. Izmed izločenih slik je bilo 8 takih, kjer ni bilo možno izračunati temperature enega, dveh ali

treh segmentov roke, medtem ko so bile preostale vrednosti uspešno zabeležene. Podatki o manjkajočih vrednostih so predstavljeni v tabeli (Tabela 6.3-1).

V podatkovno množico smo dodali teh 8 nepopolnih vzorcev in izvedli nadomeščanje manjkajočih vrednosti s pomočjo rotacijskega regresijskega gozda. Nato smo teh 8 vzorcev diagnosticirali s pomočjo nevronske mreže, ki se je učila na preostalih vzorcih. Celoten postopek smo ponovili 10-krat. Tako dobljena povprečna uspešnost klasifikacije je bila 82,5% ± 8,7%, kar je več, kot je bila povprečna uspešnost klasifikacije, ki smo jo dobili v okviru originalne raziskave (72,2% ± 1,64%), sicer na večji množici.

Domnevamo lahko, da so bile manjkajoče vrednosti tipa MCAR (teoretično bi lahko obstajala povezava med slabo pozicionirano roko med fotografiranjem in utesnitvijo živca, zato tega ne moremo trditi z gotovostjo). Vstavljanje manjkajočih vrednosti s pomočjo rotacijskega regresijskega gozda nam je omogočilo klasifikacijo 8 vzorcev, ki jih pred tem nismo mogli vključiti v raziskavo. Naknadno jih lahko dodamo učni množici in tako povečamo akumulirano znanje diagnostičnega sistema.

7 Razprava

V uvodnem poglavju smo predstavili cilje doktorske disertacije in postavili osnovno tezo. Če ponovimo:

»*Z uporabo rotacijskega regresijskega gozda kot ansambla regresijskih dreves lahko v primerjavi s klasičnimi metodami za nadomeščanje manjkajočih vrednosti v podatkovnih bazah dosežemo večjo natančnost pri določanju manjkajočih vrednosti ob zagotavljanju ohranjanja variance podatkov.*«

S tezo v mislih se lahko sedaj ozremo nazaj na opravljeno delo, vse poizkuse na katerih smo preverjali učinkovitost naše metode za nadomeščanje manjkajočih vrednosti v podatkovnih množicah, in povzamemo, kakšne rezultate smo dosegli.

Metoda za nadomeščanje manjkajočih vrednosti, ki smo jo razvili, temelji na ansamblu modelnih regresijskih dreves, ki smo ga poimenovali »rotacijski regresijski gozd«. Ta ansambel se je v podobni konfiguraciji že dobro obnesel na področju klasifikacije, zato smo predvidevali, da se bo izkazal z visoko natančnostjo tudi pri reševanju regresijskega problema numerične predikcije. Tako kot so avtorji rotacijskega gozda v njegovi osnovni implementaciji za bazni klasifikator uporabili odločitveno drevo, smo se mi odločili za regresijsko drevo, ki zaradi svoje relativne nestabilnosti omogoča izgradnjo dovolj raznolikih osnovnih regresorjev. Tip regresijskega drevesa smo določili na podlagi uvodnih preizkusov ansambla na treh, po svoji strukturi različnih, podatkovnih množicah. Z osnovnim ciljem doseganja najboljše natančnosti pri nadomeščanju manjkajočih vrednosti tipa MCAR smo tako izbrali neklesteno modelno regresijsko drevo, ki se je v povprečju izkazalo kot najnatančnejša varianta.

Poleg doseganja visoke natančnosti nadomeščanja manjkajočih vrednosti smo v primerjavi s že obstoječimi metodami enkratnega vstavljanja želeli narediti korak naprej tudi pri ohranjanju variance podatkov, kar je od ključnega pomena pri zagotavljanju nepristranskega statističnega sklepanja na podlagi morebitnih nadaljnjih statističnih analiz, opravljenih na podatkovni množici z vstavljenimi nadomestnimi vrednostmi. Pri tem se nismo želeli zanašati na temeljno sposobnost ohranjanja variance, ki jo ponuja osnovna varianta rotacijskega gozda, temveč smo razvili dodatno stohastično metodo, ki modificira napovedane vrednosti ter tako zagotavlja boljšo stopnjo ohranjanja variance. Ta metoda spreminja začetne napovedane vrednosti na podlagi nestrinjanja napovedi osnovnih regresijskih dreves ansambla, tako da izvede večje spremembe, kadar se napovedi posamezne manjkajoče vrednosti močneje razlikujejo med seboj. Oblikovali smo dve različici te metode. Pri prvi nismo želeli občutneje

spreminjati natančnosti ansambla, zato smo omejili modifikacijo vsake posamezne napovedi na ozek interval med aritmetičnim povprečjem vrednosti, ki jo za to manjkajočo vrednost napovejo osnovna regresijska drevesa, in obteženim povprečjem teh vrednosti, ki se po svoji velikosti nahaja med geometrični in harmoničnim povprečjem. Na ta način smo omilil vpliv močno odstopajočih napovedi (osamelcev). Druga, agresivnejša varianta metode modificira napovedane vrednosti znotraj intervala med obema skrajnostma standardnega odklona vseh napovedi posameznih regresijskih dreves.

Vse tri različice ansambla (osnovno različico, varianto z dodatno metodo za ohranjanje variance in varianto z agresivno metodo za ohranjanje variance) smo primerjali med seboj na različnih podatkovnih množicah. Da smo lahko realno ocenili učinkovitost našega rotacijskega regresijskega gozda, smo v primerjavo vključili tudi nekatere izmed najpogosteje uporabljanih metod za nadomeščanje manjkajočih vrednosti. Izmed tradicionalnih trivialnih metod smo uporabili vstavljanje ničel in vstavljanje povprečne vrednosti. Glavnino preostalih upoštevanih metod lahko uvrstimo med metode enkratnega vstavljanja. Tako smo v primerjavo dodali linearno regresijo, metodo najmanjših srednjih kvadratov, metodo k-najbližjih sosedov in regresijsko drevo. Za neposredno primerjavo ansambla z njegovo osnovno metodo smo vključili tudi modelno drevo. Izmed kompleksnejših metod enkratnega vstavljanja smo dodali še en ansambel, bagging, pri katerem smo kot osnovni regresor tudi uporabili modelno regresijsko drevo. Kot zadnjo smo izbrali še metodo orodja za statistično analizo SPSS, ki sodi med metode večkratnega vstavljanja.

Zaradi odločilnega vpliva, ki ga pri učinkovitosti posameznih metod za nadomeščanje manjkajočih vrednosti igra mehanizem nastanka manjkajočih podatkov, smo eksperimente izvajali posebej za vsakega izmed treh mehanizmov: popolnoma naključne manjkajoče vrednosti (MCAR), naključne manjkajoče vrednosti (MAR) in nenaključne manjkajoče vrednosti (NMAR). Natančnost metod smo ovrednotili s pomočjo metrike korena povprečne kvadratne napake (RMSE), sposobnost ohranjanja variance pa smo ocenili s primerjanjem varianc pred in po vstavljanju manjkajočih vrednosti.

Najobsežnejši del vseh poskusov smo izvedli pri nadomeščanju manjkajočih vrednosti tipa MCAR, kjer smo tudi pričakovali najboljše rezultate. Uporabili smo 14 javno dostopnih podatkovnih množic, pri čemer smo v vsako posebej naključno vstavili 7 različnih deležev manjkajočih vrednosti: 1%, 5%, 10%, 15%, 20%, 25% in 50%. Postopek smo ponovili petkrat, tako da smo dobili 490 različnih podatkovnih množic, na katerih smo uporabili vseh 12 metod za nadomeščanje manjkajočih vrednosti, tako da je končno število vseh meritev znašalo 5880. Že iz grafov, ki so ponazarjali natančnost posameznih metod, smo lahko razbrali, da sta se tako osnovna različica rotacijskega regresijskega gozda kot tudi varianta z

neagresivno metodo ohranjanja variance v veliki večini primerov izkazali bolje od preostalih metod. To so potrdili tudi statistični testi, ki so pokazali, da sta obe metodi signifikantno natančnejši od preostalih, pri čemer smo analizo izvedli kar na vseh poskusih hkrati. Različica rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance se je tudi izkazala kot statistično signifikantno natančnejša od preostalih metod, z izjemo ansambla bagging, ki je dosegel z njo primerljive rezultate. Čeprav je po svoji natančnosti zaostala za preostalima različicama našega ansambla, se je zato pri ovrednotenju sposobnosti ohranjanja variance obnesla najbolje izmed vseh metod. Na podlagi Friedmanovih rangov ji je sledila metoda orodja SPSS in nato še različica rotacijskega regresijskega gozda z neagresivno metodo ohranjanja variance. Obe najbolje uvrščeni metodi sta se izkazali za signifikantno uspešnejši od preostalih. Ker metoda orodja SPSS sodi med metode večkratnega vstavljanja in je njen osnovni cilj ohranjanje splošne variabilnosti in razmerij med atributi podatkovne množice, pa čeprav na račun natančnosti (kjer je bila med najslabšimi), na prvi pogled preseneča njena uvrstitev na šele drugo mesto. To lahko razložimo s slabo uvrstitvijo metode pri manjših deležih manjkajočih vrednosti (< 10%), medtem ko se pri višjih odstotkih obnese veliko bolje in prekosi preostale metode. Primerjava različic rotacijskega regresijskega gozda med seboj je pokazala, da varianta z agresivno metodo ohranjanja variance signifikantno bolje ohranja varianco od različice z neagresivno metodo, medtem ko je le-ta signifikantno uspešnejša od osnovne različice.

Mehanizma nastanka manjkajočih vrednosti MAR in NMAR smo simulirali na umetni podatkovni množici, s pomočjo katere smo zagotovili določeno odvisnost med posameznimi atributi. Manjkajoče vrednosti smo vstavljali samo znotraj razrednega atributa množice. 3 neodvisni atributi so nam omogočili 110 različnih neodvisnih poskusov pri nadomeščanju manjkajočih vrednosti tipa MAR, medtem ko smo na umetni podatkovni množici izvedli le 35 poskusov nadomeščanja manjkajočih vrednosti tipa NMAR. Dodatno smo nadomeščanje pri mehanizmu NMAR preverjali na izbranem atributu ene izmed javno dostopnih podatkovnih množic, s čimer smo demonstrirali vpliv porazdelitve vrednosti atributa na končne rezultate. Tako pri mehanizmu MAR kot pri NMAR so se vse tri različice rotacijskega regresijskega gozda izkazale za signifikantno natančnejše od preostalih metod. Za razliko od mehanizma MCAR, kjer se je varianta rotacijskega gozda z agresivno metodo ohranjanja variance izkazala za nekoliko manj natančno, je pri poskusih na umetni podatkovni množici, tako pri mehanizmu MAR kot tudi NMAR, prekosila preostali dve različici.

Pri uspešnosti ohranjanja variance po nadomeščanju manjkajočih vrednosti na umetni podatkovni množici izstopa metoda orodja za statistično analizo SPSS. Če se omejimo na manjkajoče vrednosti tipa

MAR, ugotovimo, da ji do neke mere sledita le regresijsko drevo in rotacijski regresijski gozd z agresivno metodo ohranjanja variance, ki sta signifikantno uspešneje ohranjala varianco od preostalih metod. Pri manjkajočih vrednostih tipa NMAR se varianta rotacijskega gozda z agresivno metodo ohranjanja variance izkaže za signifikantno boljšo od vseh preostalih metod, pri čemer pri poskusih na javno dostopni podatkovni množici prepričljivo prekosi tudi metodo orodja SPSS. Razlog za to lahko iščemo v izjemno asimetrični porazdelitvi vrednosti atributa, izbranega za vstavljanje manjkajočih vrednosti.

Če povzamemo zaključke opravljenih statističnih analiz, lahko brez težav potrdimo 1. hipotezo naše osnovne teze, ki smo jo zapisali še v prvem poglavju:

»Z uporabo nove metode, zasnovane na ansamblu raznolikih, a obenem dovolj natančnih regresijskih dreves, se v večini primerov izboljša natančnost določanja manjkajočih vrednosti v primerjavi s klasičnim metodam, kakor tudi v primerjavi z individualnim regresijskim drevesom.«

Natančnost prav vseh treh različic rotacijskega regresijskega gozda se je v povprečju izkazala kot signifikantno boljša v primerjavi s preostalimi upoštevanimi metodami.

Naša 2. hipoteza se je glasila:

»Uporaba rotacijskega regresijskega gozda za aproksimacijo manjkajočih vrednosti v večini podatkovnih baz ohranja boljšo stopnjo variance kot klasične »single-impute« metode.«

Tudi to hipotezo lahko potrdimo, če za nadomeščanje manjkajočih vrednosti uporabimo različico rotacijskega regresijskega gozda z agresivno metodo ohranjanja variance. Edina metoda, ki jo pričakovano prekosi, je metoda orodja SPSS, ki sodi med metode večkratnega vstavljanja. Pri manjkajočih vrednostih, nastalih po mehanizmu MCAR, se je tudi varianta rotacijskega gozda z neagresivno metodo ohranjanja variance obnesla bolje od preostalih metod, medtem ko je različica brez dodatne metode primerljiva s preostalimi metodami enkratnega vstavljanja. Ker se je izmed vseh treh različic rotacijskega gozda prav vsakič najbolje obnesla varianta z agresivno metodo ohranjanja variance in je tudi različica z neagresivno metodo bila vedno uspešnejša od osnovne različice, lahko potrdimo tudi 3. hipotezo:

»Z uvedbo stohastične metode, ki upošteva zanesljivosti predikcij posameznih manjkajočih vrednosti, lahko dodatno izboljšamo ohranjanja variance ob zanemarljivem vplivu na natančnost določanja manjkajočih vrednosti.«

Pri tem moramo upoštevati, da je pri manjkajočih vrednostih tipa MCAR razlika v natančnosti med različico z agresivno metodo za ohranjanje variance na eni strani in preostalima različicama rotacijskega gozda na drugi nezanemarljiva, čeprav so vse tri metode uvrščene najvišje. Razlike v natančnosti med osnovno varianto in varianto z neagresivno metodo ohranjanja variance so zanemarljive pri vseh mehanizmih.

Omeniti moramo še eno razliko med metodami enkratnega vstavljanja in metodami večkratnega vstavljanja, kot je metoda orodja SPSS. Na podlagi rezultatov opravljenih statističnih testov, še posebej Friedmanovih rangov, ne dobimo prave slike o razliki v pristopu do nadomeščanja manjkajočih vrednosti med omenjenimi metodami. Metoda orodja za statistično analizo SPSS je edina, ki pri ohranjanju variance skrbi za ohranjanje ocene negotovosti. Medtem ko se pri vseh preostalih metodah z naraščanjem deleža manjkajočih vrednosti tipa MAR in NMAR praviloma zmanjšuje varianca (saj se zmanjšuje varianca preostalih, nemanjkajočih vrednosti), poskuša metoda SPSS varianco obdržati na dovolj visoki stopnji, ki ponazarja negotovost napovedi nadomestnih vrednosti. Če bi skladno s pričakovanim postopkom večkratnega vstavljanja izvedli več ponovitev vstavljanj na isti podatkovni množici (z istimi manjkajočimi vrednostmi) in nato na tako pridobljenih popolnih podatkih opravili nadaljnjo statistično analizo, bi prišla ta razlika res do izraza.

Naknadno smo naš rotacijski regresijski gozd uporabili na praktičnem primeru podatkovne množice, pridobljene iz termografskih slik, in pokazali, da lahko z nadomeščanjem manjkajočih vrednosti v nadaljnje delo uspešno vključimo tudi vzorce podatkovne množice, ki bi jih bili drugače primorani izpustiti. Te vzorce smo nato uspešno klasificirali.

V okviru doktorske naloge smo dosegli naslednje izvirne prispevke:

- uporaba rotacijskega gozda kot ansambla regresijskih dreves za implementacijo metode nadomeščanja manjkajočih vrednosti v podatkovnih bazah,
- uvedba stohastične metode, temelječe na zanesljivosti predikcije regresijskih dreves, za izboljšanje ohranjanja variance podatkov,
- uporabniku prijazno okolje za izvedbo nadomeščanja manjkajočih vrednosti, neodvisno od podatkovne baze,
- pridobivanje novega znanja z uporabo izpopolnjenega orodja za podatkovno rudarjenje:

- podatkovno rudarjenje smo opravili na podatkovni bazi, pridobljeni z analizo termografskih slik rok pacientov s sindromom karpalnega kanala, kjer imamo opravka z manjkajočimi vrednostmi, ki jih ne moremo zanemariti.

Čeprav smo potrdili vse tri hipoteze in s tem tudi tezo doktorske disertacije, moramo izpostaviti tudi določene slabosti, oziroma dejstva, ki zgolj na podlagi opravljenih statističnih analiz niso očitna. Časovna zahtevnost naše metode je občutno višja od preostalih metod, ki ne temeljijo na ansamblu. Rotacijski gozd, ki smo ga uporabili, je v našem primeru pri vsakem vstavljanju zgradil 10 osnovnih regresijskih dreves, kar je v praktičnem primeru pomenilo, da je posamezen poskus lahko trajal več ur. Res pa je, da je postopek potrebno izvesti le enkrat, kar je ponavadi lažje, kot ponoviti manjkajoče meritve ali izpolniti anketna vprašanja. Zaradi neizpolnjenih predpogojev (porazdelitev vrednosti) ni bilo možno uporabiti parametričnih statističnih testov (t-test, ANOVA), tako da so opravljene statistične primerjave temeljile na rangiranju posameznih meritev, kar ne prikaže povsem realne slike o razlikah med metodami. Na podlagi grafov lahko ugotovimo, da so dejanske razlike med metodami relativno majhne. Zaradi predhodne normalizacija podatkov in uporabe metrike RMSE je težko oceniti, v kolikšni meri je posamezna metoda natančnejša od druge, vendar so relativne razlike pogosto manjše od nekaj odstotkov.

Če povzamemo naštete rezultate, lahko predlagamo naslednja priporočila:

- Če je le možno, se je dobro izogniti nadomeščanju s povprečnimi vrednostmi in nadomeščanju z ničlami, ter uporabiti katero izmed naprednejših metod.
- Če mehanizem nastanka manjkajočih vrednosti ni MCAR in je delež manjkajočih vrednosti relativno visok (> 5%), je edino smiselno uporabiti metodo večkratnega vstavljanja.
- Če časovna zahtevnost ni omejitev, lahko priporočimo uporabo rotacijskega regresijskega gozda za nadomeščanje manjkajočih vrednosti tipa MCAR.
- Če mehanizem nastanka manjkajočih vrednosti ni MCAR in je delež manjkajočih vrednosti relativno nizek (≤ 5%), je uporaba rotacijskega regresijskega gozda najboljša alternativa metodi večkratnega vstavljanja (vendar le, če ta ni na voljo).

8 Zaključek

Skozi sedem poglavij te disertacije smo spoznali problematiko manjkajočih vrednosti v podatkovnih množicah in načine ravnanja z njimi, od tradicionalnih, preprostih pristopov, do naprednih metod za nadomeščanje manjkajočih vrednosti, med katere sodi tudi naš rotacijski regresijski gozd. Najprej smo predstavili osnove s področja strojnega učenja in statističnih metod, potrebne za lažje razumevanje dela, ki smo ga opravili. Sledil je opis mehanizmov nastanka manjkajočih vrednosti, vzrokov in posledic, ki jih poskušamo odpraviti z nadomeščanjem manjkajočih podatkov. Nato smo predstavili rotacijski gozd, ansambel odločitvenih dreves in njegovo varianto, ki smo jo priredili za reševanje problema numerične predikcije, tako da smo odločitvena drevesa nadomestili z modelnimi regresijskimi drevesi in modelu dodali še metodo za ohranjanje variance vstavljenih vrednosti.

Uspešnost naše metode smo preverjali na dveh področjih: natančnosti napovedovanja numeričnih vrednosti in sposobnosti ohranjanja variance po vstavljanju le-teh. Rotacijski regresijski gozd smo primerjali z 9 drugimi metodami za nadomeščanje manjkajočih vrednosti, pri čemer je večina sodila v isto kategorijo ti. metod enkratnega vstavljanja, kot tudi naš ansambel. Na podlagi obsežnih empiričnih preizkusov smo potrdili v uvodu postavljene hipoteze, ki so predvidevale višjo natančnost ansambelskega pristopa k vstavljanju manjkajočih vrednosti, podobno kot se je to že pokazalo pri klasifikacijskih problemih, in boljše ohranjanje variance v primerjavi s preostalimi metodami enkratnega vstavljanja. Na tem področju je prednjačila metoda orodja za statistično analizo SPSS, kar ne preseneča, saj je predstavnica povsem drugačnega pristopa k nadomeščanju manjkajočih podatkov, predvsem iz vidika nadaljnje statistične analize na tako dobljenih popolnih podatkih.

Pridobljenih rezultatov nismo ovrednotili samo s pomočjo grafov in tabel, temveč smo opravili tudi statistične analize s pomočjo Friedmanovega in Wilcoxonovega neparametričnega testa, ki sta potrdila signifikanco naših ugotovitev.

Pomanjkljivost naše metode se skriva predvsem v njeni časovni zahtevnosti, saj ansambelski pristop zahteva gradnjo večjega števila regresorjev kot manj kompleksne metode enkratnega vstavljanja, ki predvsem po svoji natančnosti ne zaostajajo veliko za rotacijskim regresijskim gozdom. Ta relativno majhna razlika med metodami je še dodatno skrita v rezultatih neparametričnih statističnih testov, ki temeljijo na rangiranju primerjanih metod in tako ne dajo povsem realne slike o učinkovitosti. Na to in

na še nekatere druge zaključke, do katerih smo prišli med izvajanjem naše raziskave, smo opozorili v razpravi, kjer smo strnili ugotovitve v priporočila za uporabo rotacijskega regresijskega gozda.

Za zaključek naštejmo nekatere ideje, ki so se rodile med razvijanjem naše metode in izvajanjem meritev, ter jih ni bilo možno implementirati v okviru doktorske naloge in lahko predstavljajo izzive za nadaljnje delo:

- Aritmetično sredino, ki se uporablja pri povprečenju rezultatov osnovnih metod rotacijskega gozda, bi bilo mogoče smiselno zamenjati z bolj robustno oceno, npr. geometrično sredino ali obteženim povprečjem, kot smo ga uporabili pri neagresivni metodi za ohranjanje variance.
- Relativno visoka stopnja natančnosti, ki smo jo dosegli, nam dovoljuje, da naredimo metodo za ohranjanje variance še agresivnejšo, oziroma dovolimo modifikacijo osnovnih napovedi znotraj širšega intervala.
- Adaptacija rotacijskega gozda, tako da bi ansambel na podlagi informacij o atributu izbral najbolj ustrezno osnovno metodo. Že preprosta modifikacija bi omogočila uporabo odločitvenega drevesa za diskretne atribute.
- Metodo bi lahko razširili v metodo večkratnega vstavljanja.
- ...

Literatura

Acock, A. (2005). Working With Missing Values. *Journal of Marriage and Family , 67* (4), 1012-1028.

Allison, P. (2001). *Missing Data.* Thousand Oaks: Sage.

Alpaydin, E. (2004). *Introduction to Machine Learning.* Cambridge: The MIT Press.

Asuncion, A., & Newman, D. (2007). *UCI Machine Learning Repository*. (University of California, Department of Information and Computer Science) Prevzeto 7. april 2009 iz Donald Bren School of Information and Computer Sciences, University of California: http://archive.ics.uci.edu/ml/

Bennett, D. (2007). How Can I Deal With Missing Data in my Study? *Australian and New Zealand Journal of Public Health , 25* (2), 464-469.

Bo, T., Dysvik, B., & Jonassen, I. (2004). Lsimpute: Accurate Estimation of Missing Values in Microarray Data with Least Squares Methods. *Nucleic Acids Res , 32* (3:e34).

Bohen, S., Troyanskaya, O., Alter, O., Warnke, R., Botstein, D., & Brown, P. (18. februar 2003). *Variation in Gene Expression Patterns in Follicular Lymphoma and the Response to Rituximab*. Prevzeto 12. avgust 2008 iz Stanford Genomic Resources: http://genome-www.stanford.edu/rituximab/

Breiman, L. (1996). Bagging Predictors. *Machine Learning* (24), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning* (45), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* New York: Chapman & Hall.

Brock, G., Shaffer, J., Blakesley, R., Lotz, M., & Tseng, G. (2008). Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics , 9* (12).

Dasarathy, B., & Sheela, B. (1979). A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE , 67* (5), 708-713.

Freund, Y., & Schapire, R. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences , 55* (1), 119–139.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. *Machine learning : proceedings of the Thirteenth International Conference* (str. 148-156). San Francisco: Morgan Kaufmann.

Graham, J. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling , 10*, 80-100.

Graham, J., & Donaldson, S. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology , 78* (1), 119-128.

Gupta, A., & Lam, M. (1996). Estimating Missing Values Using Neural Networks. *The Journal of the Operational Research Society , 47* (2), 229-238.

Hansen, L., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 12* (10), 993-1001.

Horton, N., & Kleinman, K. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *Am Stat , 61* (1), 79-90.

Horton, N., Lipsitz, S., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician , 57* (4), 229-232.

Jamshidian, M., & Bentler, P. (1999). Using Complete Data Routines for ML Estimation of Mean and Covariance Structures with Missing Data. *Journal of Educational and Behavioral Statistics , 24* (1), 21-41.

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* New York: John Wiley & Sons.

Kenward, M., & Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research , 8* (1), 51-83.

Kononenko, I. (1997). *Strojno učenje* (1. izd.). Ljubljana: Založba FE in FRI.

Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms.* New York: John Wiley & Sons.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news , 2/3*, 18–22.

Little, R. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association , 83* (404), 1198-1202.

Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R., & Rubin, D. (2002). *Statistical Analysis with Missing Data* (2. izd.). New York: Wiley.

Little, R., & Schenker, N. (1995). Missing Data. V G. Arminger, C. Clogg, & M. Sobel (Ured.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (str. 39-75). New York: Plenum Press.

Massachusetts Institute of Technology. (2005). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice.* (G. Shakhnarovich, T. Darrell, & P. Indyk, Ured.) Cambridge: MIT Press.

Meng, X.-L. (1994). Multiple Imputation with Uncongenial Sources of Input. *Statistical Science* (9), 538-573.

Polikar, R. (2006). Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine , 6* (3), 21-45.

Quinlan, J. (1992). *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning , 1* (1), 81-106.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology , 27* (1), 85-95.

Rao, C. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya , 26*, 329 -358.

Rencher, A. (1995). *Methods of Multivariate Analysis.* New York: John Wiley & Sons.

Rodríguez, J., Kuncheva, L., & Alonso, C. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 28* (10), 1619-1630.

Rousseeuw, P. (1984). Least Median of Squares Regression. *J. Am. Statist. Assoc. , 79* (388), 871-880.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal* (4), 227–241.

Rubin, D. (1976). Inference and Missing Data. *Biometrika* (63), 581-592.

Rubin, D. (1987). *Multiple Imputation for Non response in Surveys.* New York: J. Wiley & Sons.

Schafer, J. (1997). *The Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning , 5* (2), 197-227.

Scheffer, J. (2002). Dealing with Missing Data. *Res. Lett. Inf. Math. Sci.* (3), 153-160.

Segal, M. (1988). Regression trees for censored data. *Biometrics* (44), 35–47.

Sehgal, M., Gondal, I., & Dooley, L. (2005). Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics , 21* (10), 2417-2423.

Siegel, S., & Castellan, J. J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd Edition izd.). New York: McGraw-Hill.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., in drugi. (4. september 1998). *Yeast Cell Cycle Analysis Project*. (Stanford University) Prevzeto 12. 8 2008 iz Stanford Genomic Resources: http://genome-www.stanford.edu/cellcycle/

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., in drugi. (2001). Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* (17), 520–525.

Wang, X., Li, A., Jiang, Z., & Feng, H. (2006). Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics , 7* (32).

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2. izd.). San Francisco: Morgan Kaufmann.

Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation , 8* (7), 1341-1390.

Življenjepis

Ime in priimek:	Miroslav Palfy	
Rojen:	26.2.1978, Brežice	
Šolanje:	1984-1992	Osnovna šola Brežice
	1992-1996	Gimnazija Brežice
	1996-	Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Zaposlitev:	2005-2009	mladi raziskovalec, Univerzitetni klinični center Maribor

UNIVERZA V MARIBORU

Fakulteta za elektrotehniko, računalništvo in informatiko

IZJAVA DOKTORSKEGA KANDIDATA

Podpisani-a ___Miroslav Palfy_____, vpisna številka ____95027383_____

izjavljam,

da je doktorska disertacija z naslovom ____Nadomeščanje manjkajočih vrednosti s pomočjo rotacijskega regresijskega gozda_____

- rezultat lastnega raziskovalnega dela,
- da so rezultati korektno navedeni in
- da nisem kršil-a avtorskih pravic in intelektualne lastnine drugih.

Podpis doktorskega-e kandidata-ke:

_____