Računsko zahtevne metode

Seminarska naloga

Manjkajoče vrednosti

Gregor Vavdi, Nace Vreček

7. februar 2020

Kazalo

1 Cilj naloge

Primerjanje uspešnosti razvrščanja v skupine z linearno diskriminantno analizo ob uporabi različnih metod za imputacijo manjkajočih vrednostih ter različnih mehanizmih manjkajočih vrednosti.

2 Mehanizmi manjkajočih vrednosti in generiranje podatkov

2.1 MCAR

Mehanizem povsem naključno manjkajočih podatkov. Verjetnost, da določena vrednost manjka je popolnoma neodvisna od manjkajočih vrednosti ter vrednosti pri ostalih spremenljivkah.

MCAR - Missing completly at random. Verjetnost, da določena vrednost manjka je popolnoma neodvisna od vrednosti ki manjka, kot od ostalih vrednosti (pri ostalih spremenljivkah). Delež manjkajočih vrednosti je bil enak pri vsaki spremenljivki. Za spremenljivke X_2, X_3 in X_4 sva generirala naključne števila z enakimi verjetnostmi na celotnem intervalu za vsako spremenljivko posebej.

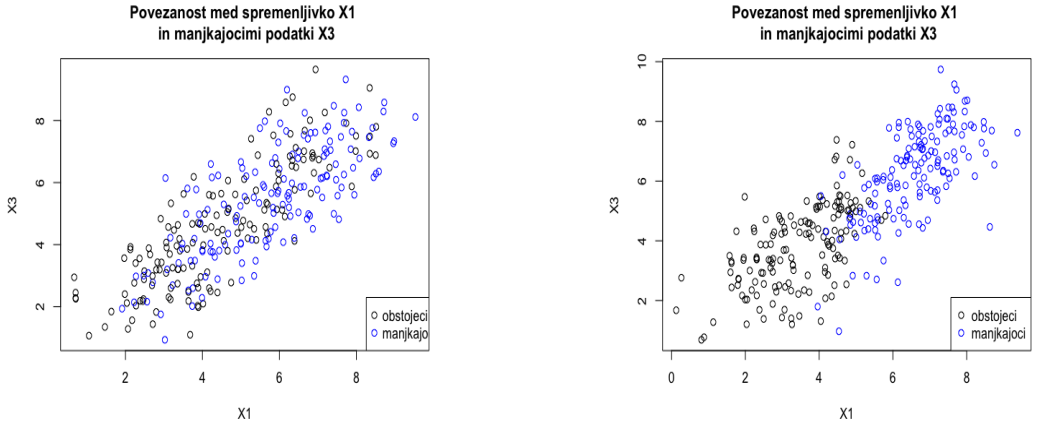
2.2 MAR

Mehanizem naključno manjkajočih podatkov. Verjetnost, da vrednost manjka je neodvisna od manjkajoče vrednosti pogojno na vrednosti pri drugih spremenljivkah.

Pri mehanizmu MAR - missing at random sva se odločila za naslednje pogojene manjkajoče vrednosti. Spremenjivka X_1 vpliva na manjkajoče vrednosti X_2, X_3 in X_4.
Povezanost med vrednostmi X_1 in manjkajočimi vrednostmi X_2 (X_3 in X_4) sva imenovala **moč** mehanizma. Ob večji povezanosti, dobimo večjo moč mehanizma. Moč sva definirala s črko m, kjer sva računala verjetnosti manjkajoče vrednosti za spremenljivko X_1:

$$p = |\frac{(X_1)^m}{(\sum_{i=1}^{n} X_1)^m}|$$

Izbirala sva med 5 različnimi vrednostmi $m : \{1, 2, 5, 8, 12\}$. Spodaj so narisani grafi za lažjo predstavo. Narisane so na vzorci N = 300 in 50% manjkajočih vrednosti, za moč: $m = 1$ in $m = 12$, pri spremenljivki X_3:



(a) Majhna moč (m = 1)

(b) Velika moč (m = 12)

Slika 1: Povezanost manjkajočih enot z obstoječimi pri X_1 in X_3

Delež NA	Skupina I	Skupina II	Skupina III
0.3	19.55	30.53	39.92
0.4	27.07	40.90	52.03
0.5	34.97	51.32	63.71
0.6	43.57	61.90	74.53

Tabela 1: Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 1

Delež NA	Skupina I	Skupina II	Skupina III
0.3	0.46	15.18	74.37
0.4	1.35	29.65	89.00
0.5	3.93	49.60	96.47
0.6	10.13	70.69	99.18

Tabela 2: Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 12

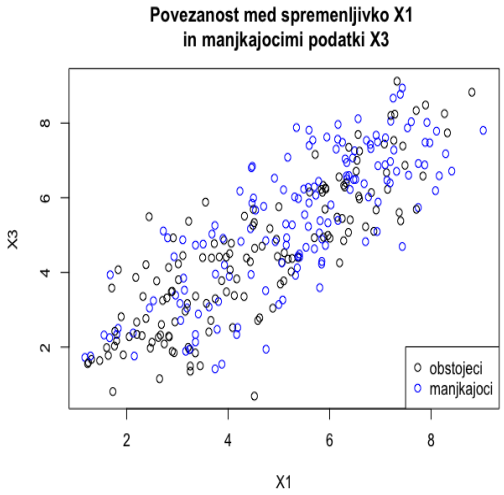
2.3 NMAR

Mehanizem nenaključno manjkajočih podatkov. Verjetnost, da vrednost manjka je odvisna od manjkajoče vrednosti (torej od spremenljivke, ki ima manjkajočo vrednost).

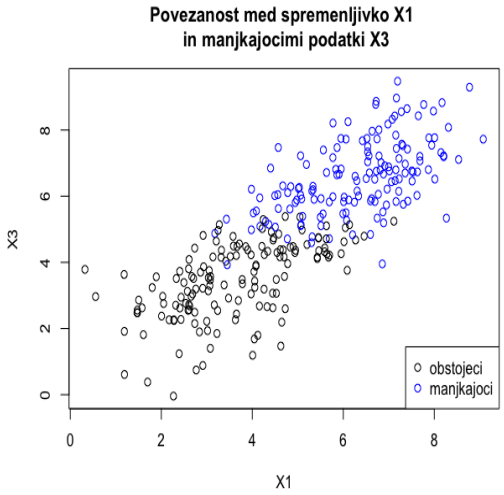
Pri mehanizmu NMAR - *Not missing at random* je spremenljivka direktno povezana s manjkajočimi vrednostmi. Tako sva manjkajoče vrednosti definirala na spremenljivki X_2, X_3 in X_4.
Povezanost med vrednostmi X_2 (X_3 in X_4) in manjkajočimi vrednostmi X_2 (X_3 in X_4) sva, podobno kot prej,imenovala **moč** mehanizma. Ob večji povezanosti, dobimo večjo moč mehanizma. Moč sva definirala s črko m, kjer sva računala verjetnosti manjkajoče vrednosti na isti spremenljivki.

$$p_i = \left|\frac{(X_i)^m}{(\sum_{i=1}^{n} X_i)^m}\right|$$

za $i = 1, 2, 3$.
Izbirala sva med 5 različnimi vrednostmi $m : \{1, 2, 5, 8, 12\}$. Spodaj so narisani grafi za lažjo predstavo. Narisane so na vzorci N = 300 in 50% manjkajočih vrednosti, za moč: $m = 1$ in $m = 12$, pri spremenljivkah X_3:



(a) Majhna moč (m = 1)



(b) Velika moč (m = 12)

Slika 2: Povezanost manjkajočih enot z obstoječimi pri X_1 in X_3

Delež NA	Skupina I	Skupina II	Skupina III
0.3	19.57	30.47	39.95
0.4	26.76	40.90	52.34
0.5	34.80	51.39	63.81
0.6	43.58	62.06	74.36

Tabela 3: Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 1

Delež NA	Skupina I	Skupina II	Skupina III
0.3	0.48	15.08	74.44
0.4	1.36	29.62	89.02
0.5	3.93	49.59	96.48
0.6	10.07	70.74	99.19

Tabela 4: Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 12

3 Metode imputacij

3.1 Analiza na podlagi popolnih enot

Analiza na podlagi popolnih enot (listwise deletion) je smiselna kadar velja MCAR mehanizem in je delež manjkajočih vrednosti majhen. V literaturi lahko zasledimo, da je taka metoda primerna v primerih, ko je delež manjkajočih vrednosti manjši od 5 % (Graham, 2009). Predpostavka o MCAR mehanizmu je močna in v praksi pogosto kršena, kar lahko privede do pristranskih ocen in manjši moči (Graham, 2009).

3.2 Multiple imputacije preko verižnih enačb (MICE)

Metoda MICE prihaja iz družine multiplih imputacij. Torej generiramo več ločenih podatkovij z imputiranimi vrednostmi (metodo imputiranja manjkajočih vrednosti določimo sami). Funkcija *mice* iz istoimenskega paketa kot metodo imputiranja uporablja *predictive mean matching*. Za vsako manjkajočo vrednost prediktorja izberemo majhen set kandidatov (cca. 10), ki nimajo manjkajočih vrednosti ter imajo napovedane vrednosti najbližje napovedani vrednosti manjkajočega prediktorja. Izmed tega seta kandidatov naključno izberemo eno vrednost ter s to vrednostjo nadomestimo manjkajočo vrednost prediktorja. Glavna predpostavka MICE metode je, da mora biti porazdelitev manjkajočih vrednosti enaka porazdelitvi opazovanih vrednosti ter, da je mehanizem manjkajočih vrednosti MCAR.

3.2.1 Koraki:

1. Vrednosti imputiramo z enostavno metodo (v najinem primeru je to *predictive mean matching*)
2. Ocenimo model za eno (vsako spremenljivko, ki ima manjkajoče vrednosti) spremenljivko, na podlagi imputiranega podatkovja iz 1. koraka.
3. Na podlagi modela iz točke 2. imputiramo nove vrednosti za manjkajoče vrednosti
4. Ponavljamo koraka 2 in 3 (za vse spremenljivke z manjkajočimi vrednostmi) dokler se porazdelitve parametrov v točki 2 ne stabilizirajo (imputacije iz 3. točke predstavljajo eno imputirano podatkovje)
5. Ponavljamo korake 1-4, da dobimo željeno število imputiranih podatkovij oziroma dosežemo konvergenco imputiranih vrednosti

3.3 Imputacije preko slučajnih gozdov

Metoda imputacij preko slučajnih gozdov je primerna tako za kvalitativne kot za kvantitavitne spremenljivke. Metoda deluje po principu napovedovanja manjkajočih vrednosti preko slučajnih gozdov, ki so bili izdelani za vsako spremenljivko (prediktor), ki ima manjkajoče vrednosti, na ne-manjkajocih vrednostih ter ta postopek ponavljamo, s tem, da vsako iteracijo posodobimo manjkajoče vrednosti z napovedmi modela, dokler ne dosežemo željene natančnosti (razlike med i in i_{-1} iteracijo) (Stekhoven & Bühlmann, 2011).

3.3.1 Psevdo algoritem:

```
1. Matrika X, velikosti nxp, zaključitveni kriterij Gama
2. Vrednosti imputiramo z enostavno metodo
3. Vektor k v katerem so spremenljivke urejene glede na delež
manjkajočih vrednosti (naraščajoč vrstni red)

while not Gama do:
    X_old_imp  ; shranimo imputirano matriko
    for s in k:
        random forest y_obs_s x_obs_s
        napovemo y_miss_s
        z napovedmi y_miss_s posodobimo matriko X_old_imp
    end for
end while
return X_old_imp
```

3.4 Metoda imputacij na podlagi najbližjih sosedov (kNN)

Metoda imputacij na podlagi najbližjih sosedov je metoda, ki temelji na podobnosti (oz. razdalji) med podatki (Beretta & Santaniello, 2016). Algoritem za ocenjevanje manjkajočih vrednosti uporabi samo popolne enote. Za oceno manjkajoče vrednosti na spremenljivki $x1$, pogledamo k vrednosti enot spremenljivke $x1$, katerih vrednosti so najbolj podobne vrednosti enote, ki ima manjkajočo vrednost. Glede optimalne določitve k ni jasnega konsenza. Tako sta Lall in Sharma (1996), kot primerno metodo ocenjevanja optimalnega števila najbližjih sosedov podala enačbo $k = \sqrt{n}$, če je $n > 100$. Kot metodo imputiranja manjkajočih vrednosti sva izbrala metodo obteženih povprečij. Uteži se izračunajo kot $e^{-d(k,x)}$, kjer d predstavlja Evklidsko razdaljo med enoto z manjkajočo vrednostjo (x) ter $k - tim$ sosedom.
Predpostavka kNN metode je, da lahko napovemo manjkajočo vrednost, na podlagi podobnih vrednosti drugih spremenljivk, ki nimajo manjkajočih vrednosti, torej, da obstaja povezanost med spremenljivko z manjkajočo enoto ter ostalimi spremeljivkami.

3.5 EM algoritem

EM-algortem temelji na dveh predpostavkah. Opazovane vrednosti so neodvisne v p razsežnem normalnem prostoru s parametri μ in Σ. Algoritem ocenjuje po metodi največjega verjetja. Druga predpostavka pravi, da podatki manjkajo naključno (MCAR) in neodvisno, vendar tako, da nikoli ne manjkajo vse komponente.

V praksi sva uporabila programsko knjižnico `norm` in njene funkcije `em.norm`, ki naredi EM - algoritem na nepopolnih podatkih za vsako skupino posebaj. Potem sva z funkcijo `getparam.nor`, kjer sva dobila parametre porazdelitve za vsako skupino posebaj. Te parametre sva uporabila za generiranje novih podatkov, za vsako skupino posebaj. Potem sva združila nepopolni set in zamenjala NA vrednosti z novimi generiranimi podatki, za vsako skupino posebaj. Nato sva združila v eno podatkovje in izvedla LDA.

4 Simulacija

Simulacijo sva razdelila na tri dele glede na tri mehanizme manjkajočih vrednosti. V vsakem delu preko razvrščanja v skupine z LDA preverjala uspešnost imputacijskih metod (za referenco sva uporabila uspešnost razvrstitve LDA modela na popolnem podatkovju). Originalno (popolno) podatkovje sva generirala iz multivariatne normalne porazdelitve ter glede na mehanizem manjkajočih vrednosti (MAR, MCAR, NMAR) uvedla manjkajoče vrednosti (katerih delež se je spreminjal tekom simulacije). Kot končno mero uspešnosti sva uporabila delež pravilno razvrščenih enot na novem podatkovju, brez manjkajočih vrednosti.

4.1 Fiksni dejavniki

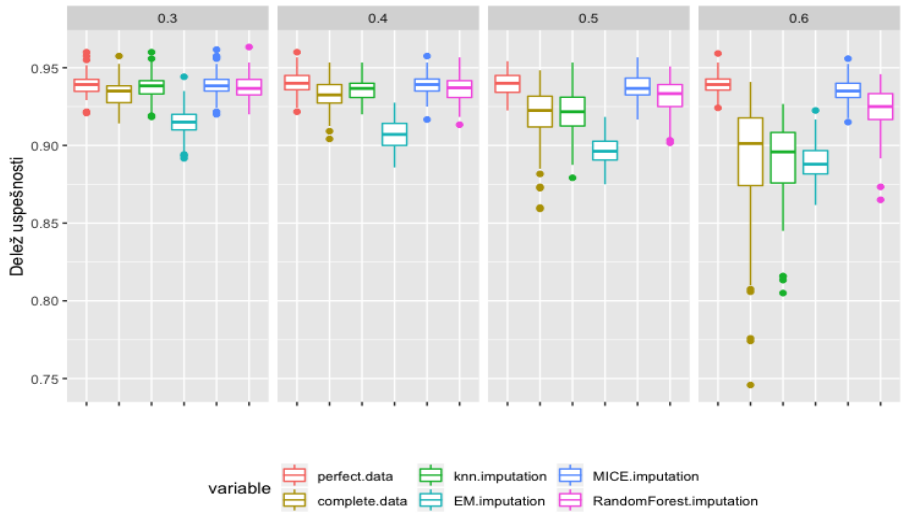
- Število spremenljivk; 4
- Število skupin; 3
- Velikost skupin; 100, 100, 100
- Povprečja skupin; 3, 5, 7
- Varianca spremenljivk; 1
- Korelacija med spremenljivkami; 0.3
- Porazdelitev spremenljivk; multivariatna normalna
- Število spremenljivk z manjkajočimi vrednostmi; 3

4.2 Variabilni dejavniki

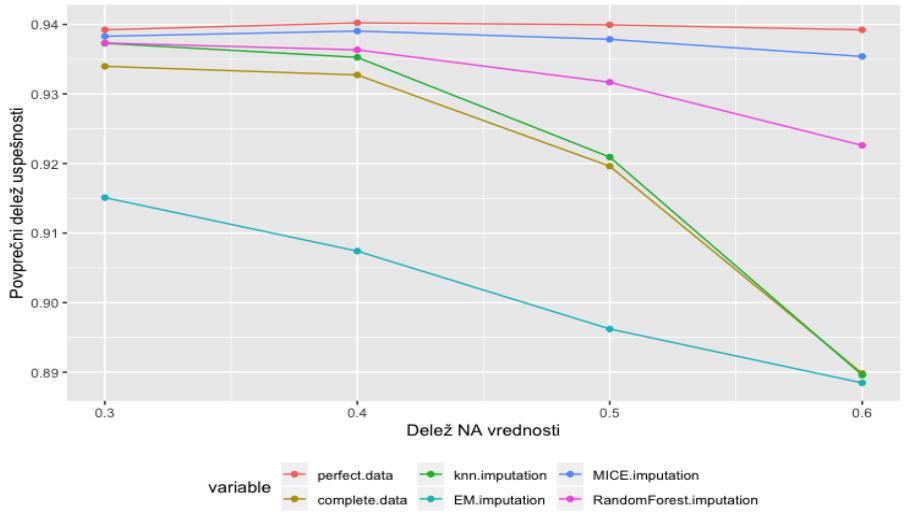
- Mehanizem manjkajočih vrednosti; MCAR, MAR, NMAR
- Moč mehanizma manjkajočih vrednosti (pri MAR in NMAR); 1, 2, 5, 8, 12
- Metoda imputacije manjkajočih vrednosti
- Delež manjkajočih vrednosti; 0.3, 0.4, 0.5, 0.6

5 Rezultati

5.1 Rezultati MCAR

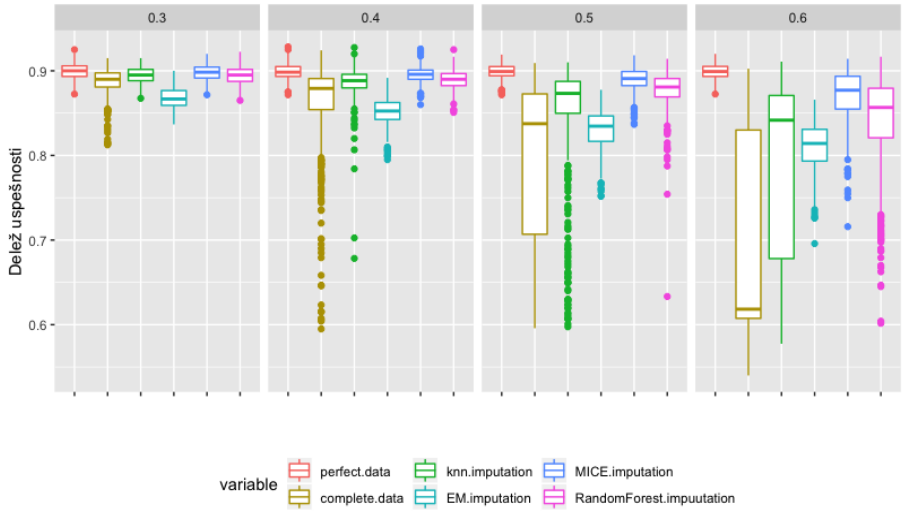


Slika 3: Porazdelitev...

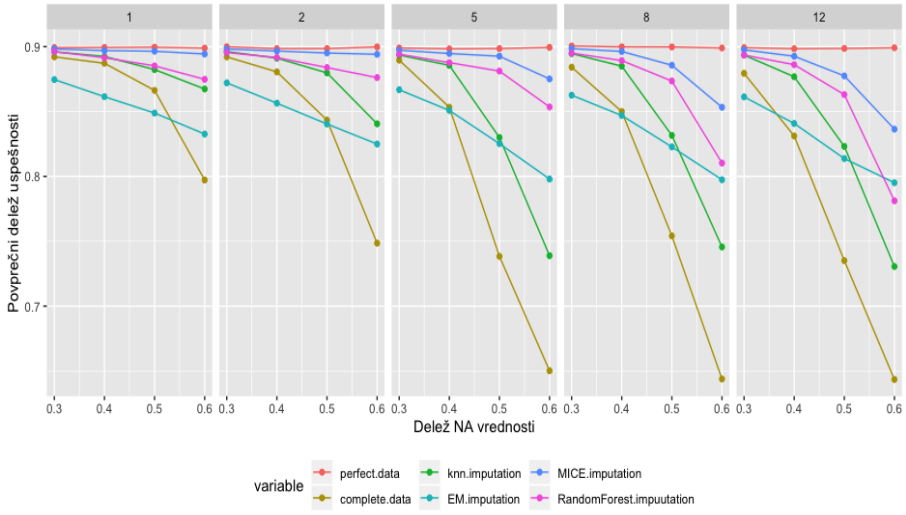


Slika 4: Povprečje uspešnosti pri mehanizmu MCAR za različne deleže manjkajočih vrednosti

5.2 Rezultati MAR

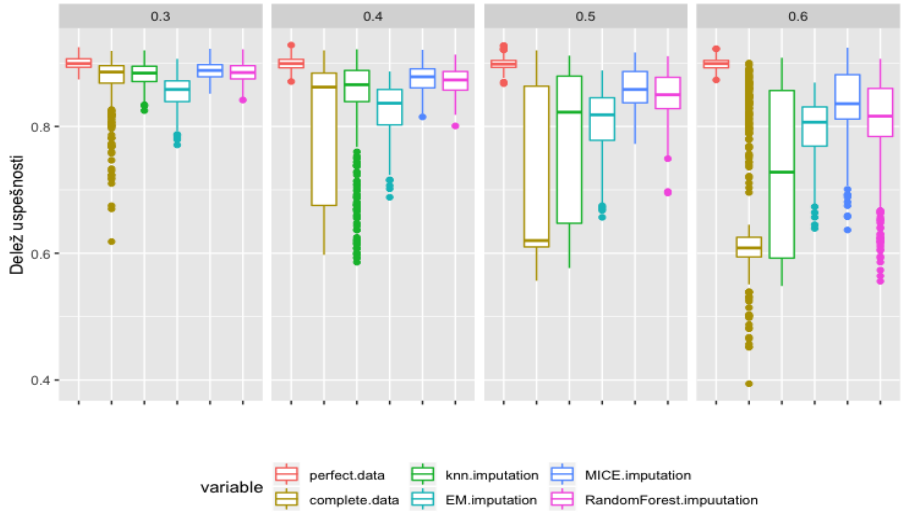


Slika 5: Porazdelitev...

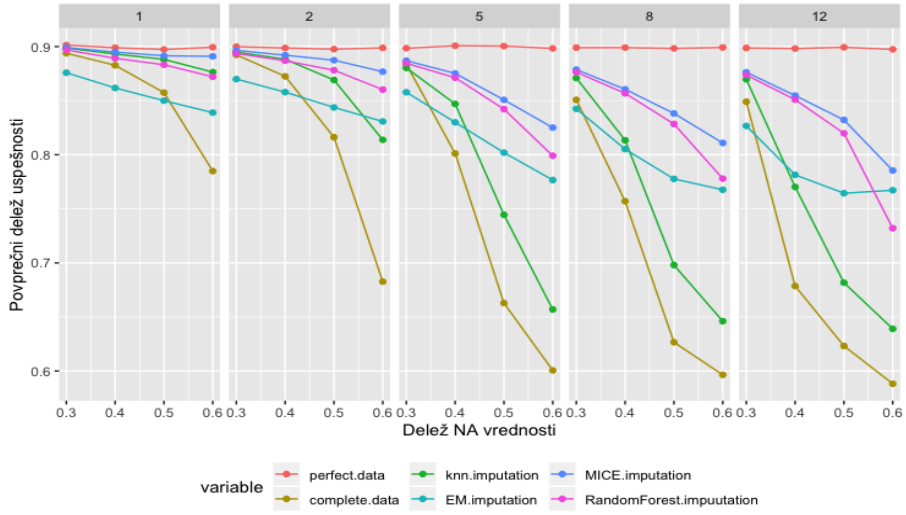


Slika 6: Povprečje uspešnosti pri mehanizmu MAR za različne deleže manjkajočih vrednosti in moči

5.3 Rezultati NMAR



Slika 7: ...



Slika 8: Povprečje uspešnosti pri mehanizmu NMAR za različne deleže manjkajočih vrednosti in moči

6 Zaključek

Ne pozabi na priporočila.

Literatura

[1] Beretta, L., Santaniello, A. (2016). *Nearest neighbor imputation algorithms: a critical evaluation.* BMC Med Inform Decis Mak. 2016;16(Suppl 3):74. doi: 10.1186/s12911-016-0318-z.

[2] Graham, J. W. (2009). *Missing data analysis: Making it work in the real world.* Annual Review of Psychology, 60(1), 549–576. doi:10.1146/annurev.psych.58.110405.085530

[3] Lall, U. & Sharma, A. (1996). *A nearest-neighbor bootstrap for resampling hydrologic time series.* Water Resources Research 32 (3), 679–693.

[4] Stekhoven, D. J., & Bühlmann, P. (2011). *MissForest: non-parametric missing value imputation for mixed-type data.* Bioinformatics, 28(1):112-118.

[5] Geoffrey J. McLachlan, Thriyambakam Krishnan, The EM Algorithm and Extensions, Wiley Series in Probability and Statistics, 1997.