

Domaca naloga 4

Neža Kržan, Tom Rupnik Medjedovič

1 Cilj naloge

Analizirali bomo podatke o diabetesu Nacionalnega inštituta za diabetes, prebavne in ledvične bolezni, kateri vsebujejo več medicinskih napovednih spremenljivk in eno ciljno spremenljivko. Podatki imajo veliko manjkajočih enot (več kot 30%), zato bova v nalogi definirala oz. ocenila, kakšen je mehanizem manjkajočih vrednosti (po Rubinu) in obravnavala manjkajoče vrednosti po treh različnih metodah.

Zanima naju torej mehanizem manjkajočih vrednosti (zakaj in kako manjkajo podatki v podatkovnem nizu), ker ti mehanizmi vplivajo na izbiro ustreznih metod za obravnavo manjkajočih vrednosti, kar pa bova delala v drugem delu naloge.

2 Podatki

Izbrala sva si zdravstvene podatke žensk indijanskega plemena Pima, ki so starejše od 21 let. Glavni cilj tega niza je napovedati, ali ima posameznica diabetes, na podlagi različnih medicinskih spremenljivk:

- **Pregnancy** (število nosečnosti): Število nosečnosti, ki jih je imela ženska.
- **Glucose** (glukoza v krvi): Raven glukoze v krvi po 2 urah oralnega glukoznega testa.
- **BloodPressure** (krvni tlak): Krvni tlak (v mmHg).
- **SkinThickness** (debelost tkivo): Debelina kožnega gubca (v mm) na tricepsu, merjeno za testiranje telesne maščobe.
- **Insulin** (inzulin): Raven inzulina (v $\mu\text{U/ml}$) v krvi.
- **BMI** (Body Mass Index): Indeks telesne mase (BMI).
- **DiabetesPedigree** (genetska nagnjenost): Indikator, ki kaže, koliko je posameznica nagnjena k razvoju diabetesa na podlagi dednosti.
- **Age** (starost): Starost posameznice v letih.
- **Class** (diabetes diagnoza): Ciljna spremenljivka, ki označuje, ali ima posameznica diabetes (1) ali ne (0). To je binarna spremenljivka, ki jo želimo napovedati na podlagi drugih spremenljivk.

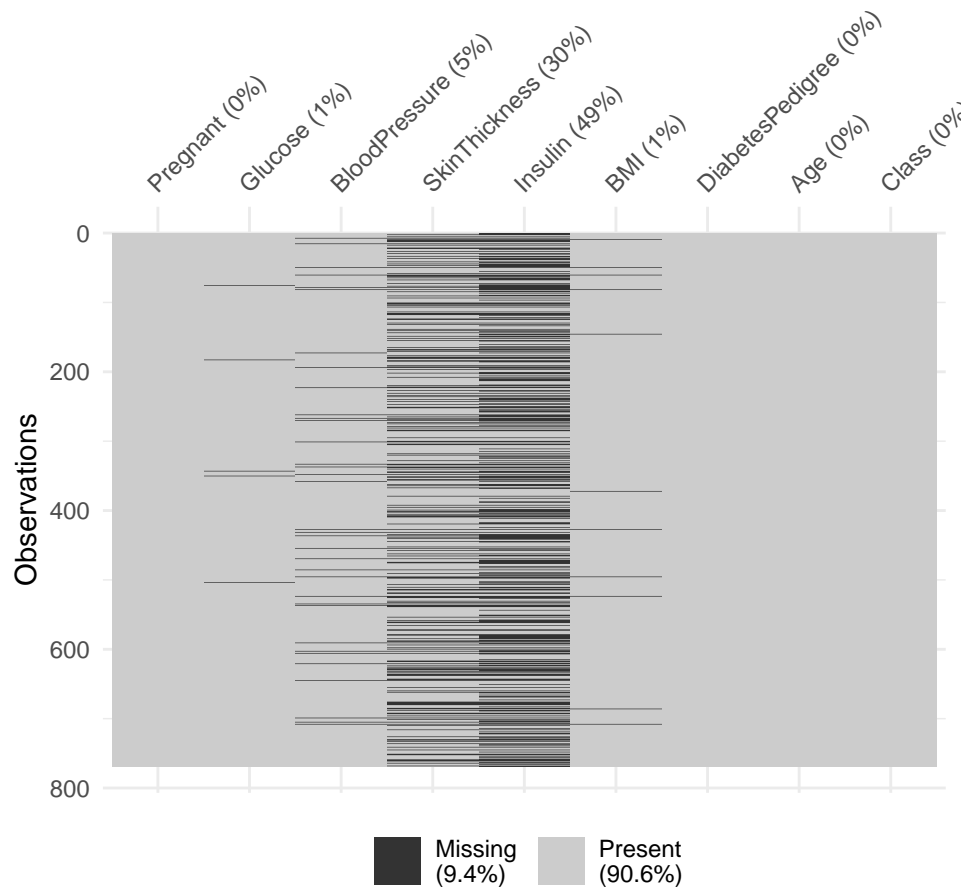
Poglejmo si osnovne statistike podatkov, iz katerih vidimo nekatere izjemne vrednosti, kot so npr. število nosečnosti 17, vrednost inzulina 846, debelost kože 99, ... Ker naju zanima analiza manjkajočih vrednosti, podatkov podrobneje ne bova analizirala.

Tabela 1: Opisne statistike podatkov.

	n	mean	sd	median	min	max	se
Pregnancy	768	3.85	3.37	3.00	0.00	17.00	0.12
Glucose	763	121.69	30.54	117.00	44.00	199.00	1.11
BloodPressure	733	72.41	12.38	72.00	24.00	122.00	0.46
SkinThickness	541	29.15	10.48	29.00	7.00	99.00	0.45
Insulin	394	155.55	118.78	125.00	14.00	846.00	5.98
BMI	757	32.46	6.92	32.30	18.20	67.10	0.25
DiabetesPedigree	768	0.47	0.33	0.37	0.08	2.42	0.01
Age	768	33.24	11.76	29.00	21.00	81.00	0.42

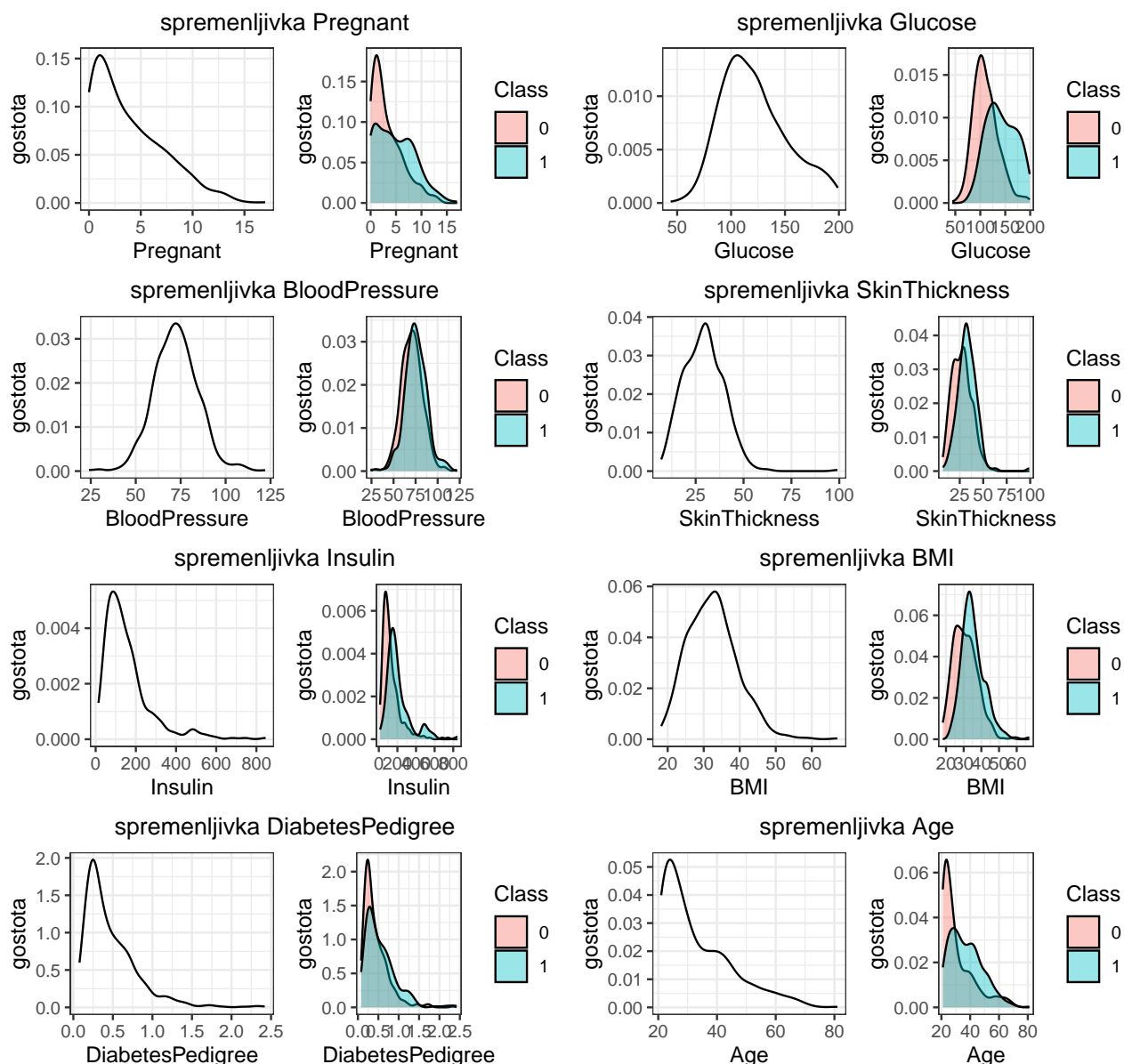
Class	768	0.35	0.48	0.00	0.00	1.00	0.02
-------	-----	------	------	------	------	------	------

Podatkovni niz vsebuje 768 primerov in torej vsak primer vključuje 8 merjenih spremenljivk in diagnozo(**Class**). Podatki vsebujejo veliko količino manjkajočih vrednosti, kar lahko vidimo na spodnjem grafu.



Slika 1: Odstotek manjkajočih vrednosti pri posamezni spremenljivki in vizualizacija.

Za lažjo predstavo si pogledjmo grafe spremenljivk v naših podatkih, iz katerih vidimo, da je večina spremenljivk asimetričnih v desno.



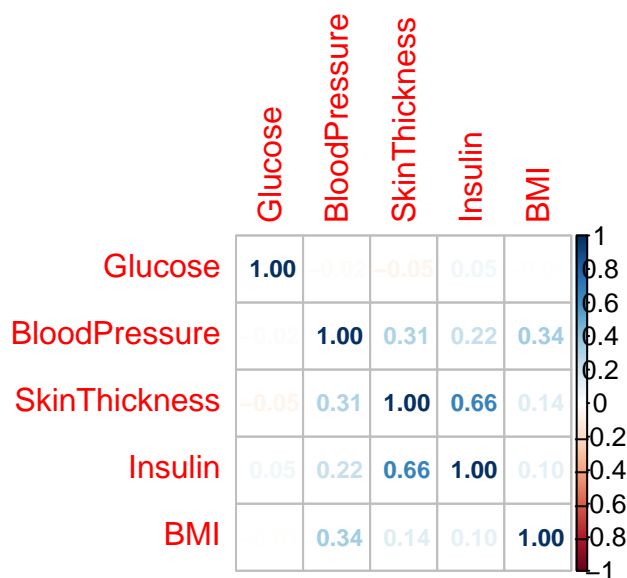
Slika 2: Porazdelitev spremenljivke v podatkovnem nizu, kjer je Class ciljna spremenljivka, ki označuje, ali ima posameznika diabetes (1) ali ne (0).

3 Mehanizem manjkajočih vrednosti

Za začetek si pogledjmo korelacije med manjkajočimi vrednostmi na spodnji korelacijski matriki. Opazimo nekaj pozitivnih korelacij, kar pomeni, da če imamo manjkajočo vrednost pri eni spremenljivki, obstaja večja verjetnost, da bo tudi pri drugi.

Torej če imamo manjkajočo vrednost pri spremenljivki **Insulin**, potem obstaja večja verjetnost, da bomo imeli manjkajočo vrednost tudi pri spremenljivki **SkinThickness** (tudi pri **BloodPressure**). Če imamo manjkajočo vrednost pri spremenljivki **SkinThickness**, obstaja večja verjetnost, da bomo imeli manjkajočo vrednost še pri **BloodPressure**. Če pa bomo imeli manjkajočo vrednost pri spremenljivki **BMI**, potem obstaja večja verjetnost, da bomo imeli manjkajočo vrednost tudi pri spremenljivki **BloodPressure**. Iz tega bi lahko sklepali, da bo t-test, ki ga uporabljamo za MAR testiranje (*missing at random*), pokazal, da odsotnost podatkov pri spremenljivkah **SkinThickness**, **Insulin**, **BloodPressure** in **BMI** ni naključno, saj vidimo, da je odsotnost

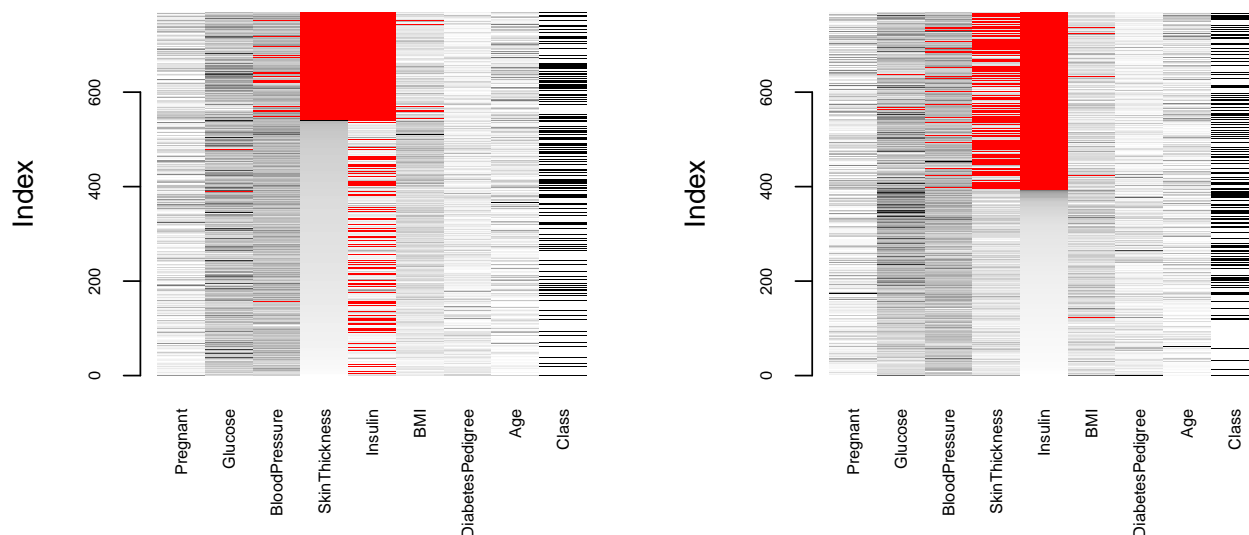
podatkov pri določenih spremenljivkah povezano.



Slika 3: Korelacije med manjšimi vrednostmi.

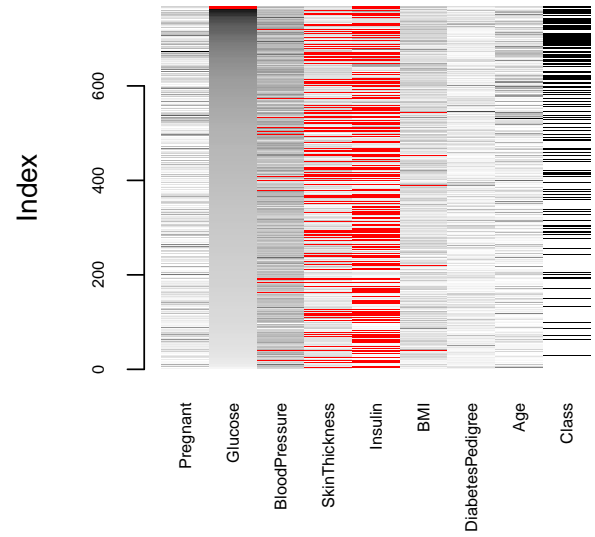
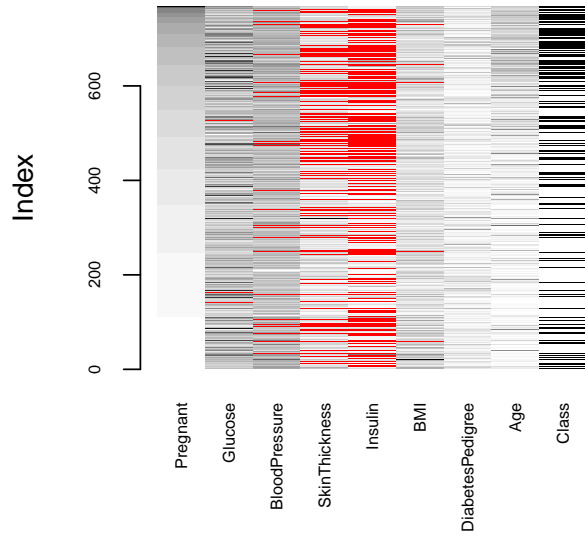
Poglejmo si še vzorce manjših vrednosti.

Najprej si pogledajmo glede razvrstitev spremenljivk, kjer imamo največ manjših vrednosti, torej spremenljivki **SkinThickness** in **Insulin**. Iz vzorov na spodnjih grafih bi lahko rekli, da če imamo vrednost pri spremenljivki **Insulin**, jo imamo tudi pri drugih spremenljivkah (le dve izjemi), če pa nimamo vrednosti pri spremenljivki **SkinThickness**, jo nimamo tudi pri spremenljivki **Insulin** in obstaja velika verjetnost, da je ne bomo imeli tudi pri spremenljivki **BloodPressure**.

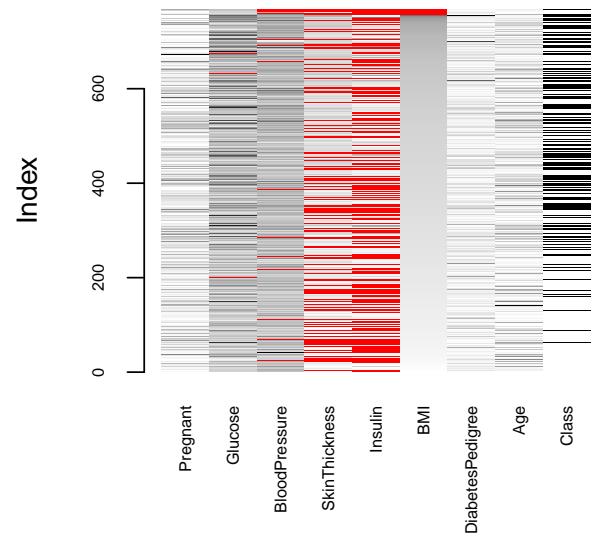
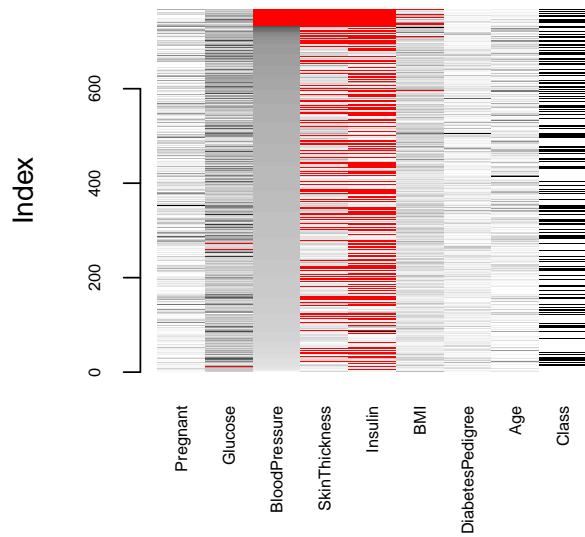


Ker je večina manjših vrednosti pri spremenljivkah **SkinThickness** in **Insulin**, nas sicer zanimajo predvsem razvrstitve glede na druge spremenljivke, zato so to tudi oglejmo.

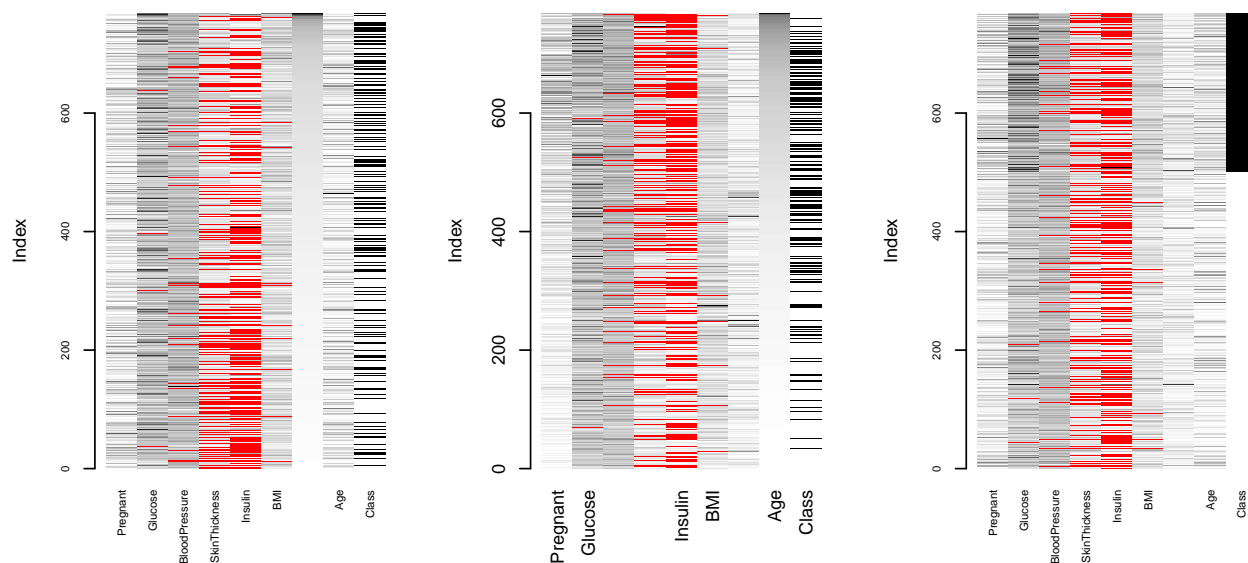
Iz spodnjega grafa imamo občutek, da obstaja večja verjetnost manjšega podatka pri spremenljivkah **Insulin** in **SkinThickness** ob določenem številu nosečnosti pri ženskah. Pri razvrstitvi glede na spremenljivko **Glucose** ne zaznamo nekega vzorca.



Iz spodnjega grafa se nam zdi, da če imamo manjšo vrednost pri spremenljivki **BloodPressure**, potem obstaja velika verjetnost, da bomo imeli manjša vrednosti tudi pri spremenljivkah **SkinThickness** in **Insulin**. Prav tako obstaja večja verjetnost manjših vrednosti pri teh dveh spremenljivkah in še pri spremenljivki **BloodPressure**, če imamo manjšo vrednost pri spremenljivki **BMI**.



Na spodnjih treh grafih pa lahko vidimo, da je pri manjših vrednostih spremenljivke **DiabetesPedigree** in višji starosti več manjših vrednosti spremenljivke **Insulin**.



V nadaljevanju sva izvedla več t-testov za preverjanje ali je odsotnost podatkov pri določenih spremenljivkah MAR(*missing at random*) ali ne. S testom sva testirala ali je povprečje posamezne spremenljivke v obeh skupinah (1. skupina so podatki, ki imajo vrednost druge spremenljivke, 2. skupina pa podatki, ki nimajo vrednost neke druge spremenljivke) enako, primer ničelne hipoteze, ki sva jo testirala

H_0 : Povprečni sprem. BMI v obeh skupinah (tistih, ki imajo podatke za Insulin in tistih, ki nimajo podatkov za Insulin) sta enaki.

S pomočjo testov sva ugotovila, da manjkajoči podatki pri spremenljivki **Insulin** niso naključni, saj rezultati testov nakazujejo, da so manjkajoče vrednosti povezane s številom nosečnosti, da obstaja večja verjetnost manjkajočega podatka pri starejših ženskah, pri tistih z višjim krvnim tlakom (**BloodPressure**) in nižjimi vrednostmi spremenljivke **DiabetesPedigree**. Manjkajoči podatki pri spremenljivki so verjetno MAR(*missing at random*) ali celo NMAR(*not missing at random*), ker so torej odvisne od drugih spremenljivk (ki načeloma ne manjkajo) in od manjkajočih vrednosti.

Prav tako je odsotnost podatkov pri spremenljivki **SkinThickness** povezana z vrednostmi pri spremenljivki **Pregnant** in višjim krvnim tlakom (sprem. **BloodPressure**) - povprečne vrednosti krvnega tlaka so v skupini z manjkajočimi podatki **SkinThickness** višje. Poleg tega so manjkajoči podatki **SkinThickness** povezani še z višjo povprečno starostjo (**Age**) in nižjimi vrednostmi **DiabetesPedigree**. Torej odsotnost podatkov v spremenljivki morda je MAR(*missing at random*) ali celo NMAR(*not missing at random*), ker so torej odvisne od drugih spremenljivk (ki načeloma ne manjkajo) in od manjkajočih vrednosti.

Pri manjkajočih vrednostih spremenljivke smo testirali, ali na manjkajoče vrednosti kaj vpliva prisotnost/odsotnost podatko spremenljivk **BMI** in **SkinThickness** - glede na zgornjo korelacijski matriko. Ugotovila sva, da so manjkajoči podatki za sprem. **BloodPressure** verjetno MCAR (Missing Completely At Random), saj prisotnost manjkajočih podatkov ni povezana z vrednostmi **BMI** ali **SkinThickness**.

Prav tako sva pri manjkajočih vrednostih spremenljivke **Glucose** ugotovila, da so verjetno MCAR (Missing Completely At Random), saj prisotnost manjkajočih podatkov ni povezana z vrednostmi nobene druge spremenljivke.

Poleg tega sva prišla do zaključka, da je verjetnost za manjkajočo vrednost pri posamezni spremenljivki (z izjemo sprem. **BMI**) večja pri ženskah, ki imajo diabetes, kot pri tistih, ki ga nimajo.

S pomočjo χ^2 -testa, ki preučuje ali obstaja statistično značilna povezava med manjkajočimi vrednostmi v posamezni spremenljivki in razredi v spremenljivki **Class**, sva ugotovila, da manjkajoče vrednosti pri spremenljivkah niso statistično povezane z razredom **Class**. Razred, ki označuje prisotnost ali odsotnost sladkorne bolezni, torej ne vpliva na to, ali manjkajo podatki za spremenljivke.

4 Logistična regresija

S pomočjo logistične regresije (`glm()`) bova na podlagi različnih medicinskih spremenljivk (meritev) poskusila napovedati, ali ima posameznica diabetes. Pri tem bova uporabila različne metode za obravnavo manjkajočih vrednosti. Najprej bova preverila kakšne vrednosti koeficientov dobimo v primeru, da uporabimo podate takšne kakršni so (originalne). Nato bova s pomočjo metode `listwise deletion` ohranila le tiste statistične enote, ki imajo zabeležene podatke vseh spremenljivk. Manjkajoče vrednosti (NA) bova z metodo *odločitvenih dreves* (`random forest`) izračunala oz. “zapolnila” ter naredila logistično regresijo. Za konec pa bova preverila kakšne rezultate dobimo v primeru uporabe metode *multiple imputacije preko verižnih enačb* (MICE).

4.1 Originalni (prvotni podatki)

Najprej preverimo kakšne vrednosti koeficientov in pripadajočih 95% intervalov zaupanja dobimo v primeru, da za modeliranje uporabimo originalne (nespremenjene podatke). Pri izvajanju tega sva ugotovila, da funkcija za logistično regresijo `glm()` ne sprejema manjkajočih vrednosti in že sama po sebi izvaja metodo `listwise deletion`.

Tabela 2: Osnovne opisne statistike podatkov.

	listwise deletion				izvirni podatki			
	n	mean	sd	se	n	mean	sd	se
Class*	392	1.33	0.47	0.02	768	0.35	0.48	0.02
Pregnant	392	3.30	3.21	0.16	768	3.85	3.37	0.12
Glucose	392	122.63	30.86	1.56	763	121.69	30.54	1.11
BloodPressure	392	70.66	12.50	0.63	733	72.41	12.38	0.46
SkinThickness	392	29.15	10.52	0.53	541	29.15	10.48	0.45
Insulin	392	156.06	118.84	6.00	394	155.55	118.78	5.98
BMI	392	33.09	7.03	0.35	757	32.46	6.92	0.25
DiabetesPedigree	392	0.52	0.35	0.02	768	0.47	0.33	0.01
Age	392	30.86	10.20	0.52	768	33.24	11.76	0.42

Če natančneje pogledamo zgornjo tabelo opisnih statistik lahko opazimo, da so razlike med statistikami, ki smo jih naredili na začetku. Funkcija `glm()` sama po sebi odstrani vse vrstice, ki vsebujejo vsaj eno NA vrednost. To je enako kot bi naredili *listwise deletion*. To lahko opazimi tudi iz stolpca, ki prikazuje število enot, na katerih so izračunane opisne statistike (vse vrednosti so enake 392).

Ker bi, ne glede na to katero preprosto metodo bi uporabila za obravnavo manjkajočih vrednosti (*listwise deletion* ali *pairwise deletion*), bili rezultati identični zgornjim, tema dvema metodama ne bova posvečala velike pozornosti, bova pa prikazala rezultate metode kot eno izmed možnosti.

4.2 Vstavljanje srednje vrednosti

Pri tej metodi na mesto manjkajočih vrednosti (NA) vstavimo srednjo vrednost spremenljivke (povprečje) izračunano na podatkih, ki jih imamo na voljo. Za metodo *vstavljanje srednje vrednosti* sva se odločila, saj zaradi vstavitve le ene vrednosti na vsa manjkajoča mesta v spremenljivki, pričakujemo da bo slabše delovala. Vendar naju vseeno zanima koliko se bo zares razlikovala od preostalih metod. Gre sicer za dokaj preprosto metodo, ki ni računsko zahtevna in manjkajoče vrednosti nadomestimo na dokaj lahek način.

Tabela 3: Opisne statistike podatkov.

	vstavljanje srednje vrednosti				izvirni podatki			
	n	mean	sd	se	n	mean	sd	se

Class*	768	1.35	0.48	0.02	768	0.35	0.48	0.02
Pregnant	768	3.85	3.37	0.12	768	3.85	3.37	0.12
Glucose	768	121.69	30.44	1.10	763	121.69	30.54	1.11
BloodPressure	768	72.41	12.10	0.44	733	72.41	12.38	0.46
SkinThickness	768	29.15	8.79	0.32	541	29.15	10.48	0.45
Insulin	768	155.55	85.02	3.07	394	155.55	118.78	5.98
BMI	768	32.46	6.88	0.25	757	32.46	6.92	0.25
DiabetesPedigree	768	0.47	0.33	0.01	768	0.47	0.33	0.01
Age	768	33.24	11.76	0.42	768	33.24	11.76	0.42

Iz tabela lahko razberemo, da smo res nadomestili vse manjkajoče vrednosti (število vrednosti v vsaki spremenljivki je enako 768). Prav tako se povprečja na novih in originalnih podatkih ujemata, kar je pričakovano, saj smo ravno povprečje vrednosti spremenljivke uporabili za nadomestitev manjkajočih vrednosti. Poleg spremembe števila vrednosti v spremenljivki, opazimo tudi spremembo standardnega odklona pri spremenljivkah z manjkajočimi vrednostmi (*SkinThickness*, *Insulin*,...)

4.3 Odločitvena drevesa

Kot naslednjo metodo, s katero bova nadomestila manjkajoče vrednosti v podatkih, sva si izbrala metodo odločitvenih dreves. Gre za metodo, ki dobro deluje na velikih podatkih, je robustna za nelinearnost, dobro deluje tudi v primeru osamelcev in jo lahko uporabimo tako na številskih kot tudi kategoričnih spremenljivkah. Občutek imava, glede na najine podatke, da bo to dobra metoda, ker imava velik nabor podatkov, večino številске spremenljivke in nekaj ekstremnih vrednosti, ki so osamelci(lahko vidimo v tabeli opisnih statistik podatkov na začetku). Odločila sva se za deterministični pristop vstavljanja manjkajočih vrednosti, torej, da vstavljava le napovedi(brez slučajne napake modela).

Tabela 4: Opisne statistike podatkov.

	odločitvena drevesa				izvirni podatki			
	n	mean	sd	se	n	mean	sd	se
Class*	768	1.35	0.48	0.02	768	0.35	0.48	0.02
Pregnant	768	3.85	3.37	0.12	768	3.85	3.37	0.12
Glucose	768	121.64	30.50	1.10	763	121.69	30.54	1.11
BloodPressure	768	72.35	12.15	0.44	733	72.41	12.38	0.46
SkinThickness	768	28.86	9.43	0.34	541	29.15	10.48	0.45
Insulin	768	154.00	99.83	3.60	394	155.55	118.78	5.98
BMI	768	32.40	6.91	0.25	757	32.46	6.92	0.25
DiabetesPedigree	768	0.47	0.33	0.01	768	0.47	0.33	0.01
Age	768	33.24	11.76	0.42	768	33.24	11.76	0.42

Tokrat lahko vidimo, da smo z izračunom nadomestili manjkajoče vrednosti (število vrednosti v vsakem stolpcu (spremenljivki) je enako 768, tako kot na izvirnih podatkih). Pričakovano so se spremenile vrednosti opisnih spremenljivk, predvsem pa opazimo razliko oz. zmanjšanje standardnega odklona (**sd**) pri vseh spremenljivkah, opaziva tudi, da so zmanjšanja/večanja opisnih statistik bolj izrazita pri spremenljivkah, ki so imele v izvirnih podatkih več manjkajočih vrednosti(npr. *Insulin*), kar je seveda logično, saj moramo vstaviti več manjkajočih vrednosti.

4.4 Multiple(stohastične) imputacije preko verižnih enačb

Najprej si oglejmo tabelo manjkajočih vrednosti. Desna stran predstavlja število spremenljivk z manjkajočimi vrednostmi, leva stran pomeni število enot, ter spodaj imamo število enot z manjkajočimi vrednostmi.

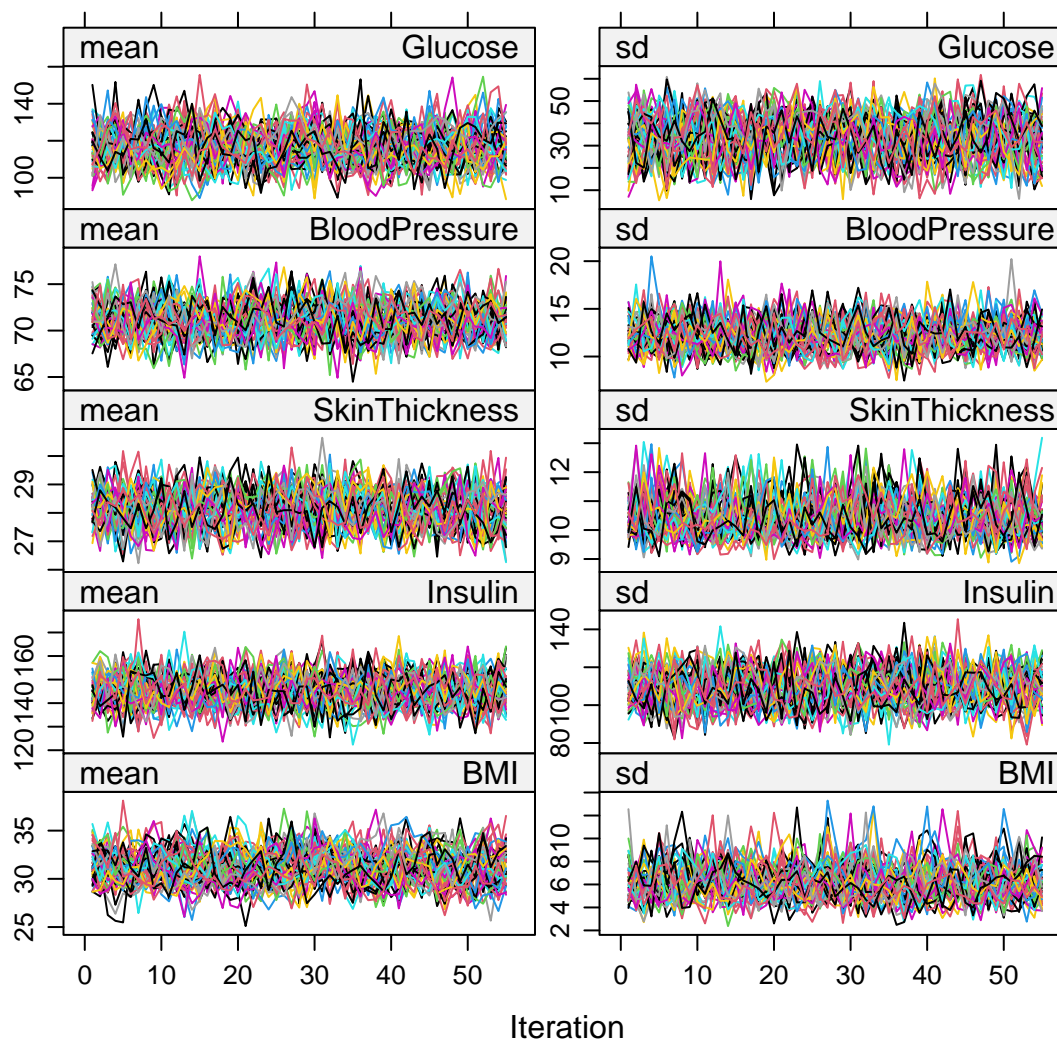
	Pregnant	DiabetesPedigree	Age	Class	Glucose	BMI	BloodPressure	SkinThickness	Insulin	
392										0
140										1
192										2
2										2
26										3
1										1
1										2
2										3
7										4
1										1
4										2
	0	0	0	0	5	11	35	22	73	652

Slika 4: Tabela manjkajočih vrednosti.

Enotam z manjkajočimi vrednostmi najpogosteje manjkajo vrednosti spremenljivk `Insulin(1)`, `SkinThickness(2)` ali `BloodPressure(3)`, kar smo ugotovili že pri mehanizmu manjkajočih vrednosti in analizi podatkov. Prav tako skoraj v vseh primerih je vrednost spremenljivke `Insulin` manjkajoča vrednost (le dve enoti v podatkih imata manjkajoče vrednosti in ta ni pri spremenljivki `Insulin`, ampak pri sprem. `BMI` in `Glucose`).

Za izračun manjkajočih vrednosti sva uporabila funkcijo `mice()`, kjer sva za vrednosti prametrov izbrala `m = 50` (pravilo palca pravi da izberemo `m` glede na % manjkajočih vrednosti - imamo 51.04% manjkajočih enot) in `maxit = 55` (nekoliko višja vrednost, da vrednosti zagotovo skonvergirajo).

Poglejmo si grafe povprečja in standardnega odklona za spremenljivke z manjkajočimi vrednostmi, da se prepričamo ali vrednosti res skonvergirajo.



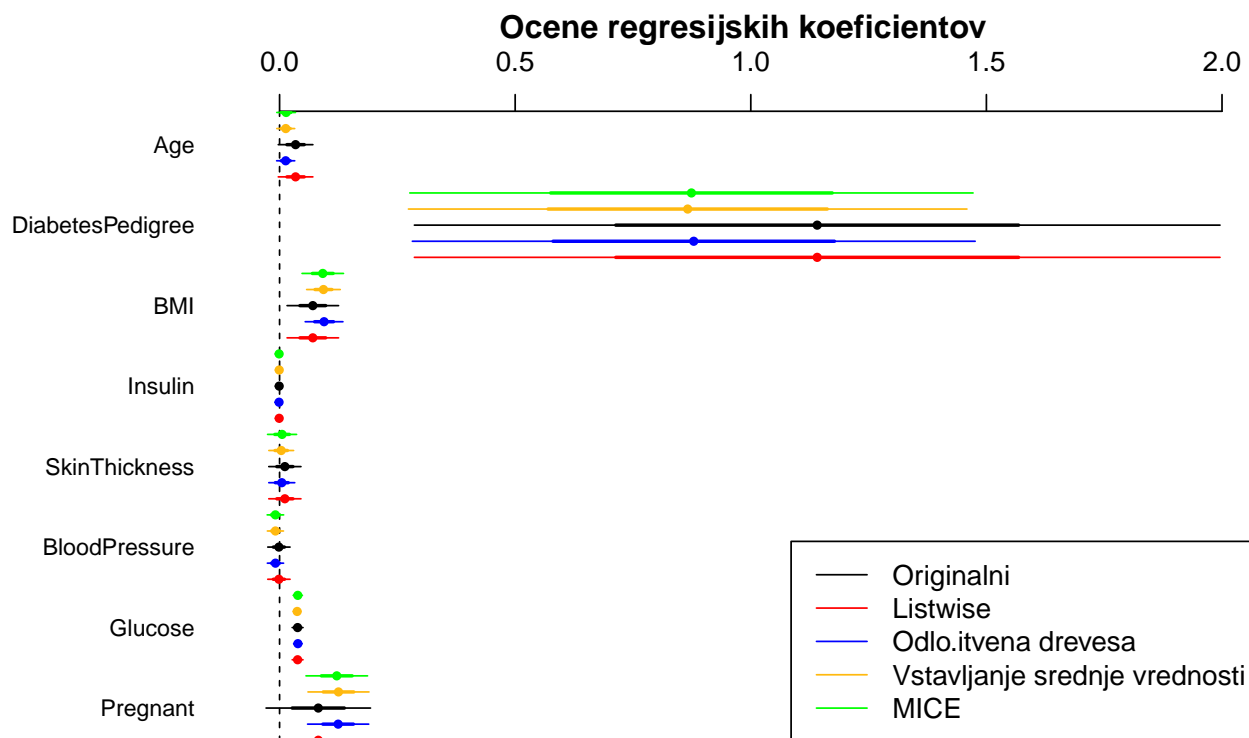
Slika 5: Grafi povprečja in standardnega odklona spremenljivk z manjkajočimi vrednostmi.

Glede na zgornje grafe, bi lahko rekli, da smo izbrali prave vrednosti parametrov (vrednosti pri vseh spremenljivkah z iteracijami ustalijo - gledamo grafe od leve proti desni).

5 Primerjava regresijskih koeficientov in intervalov zaupanja

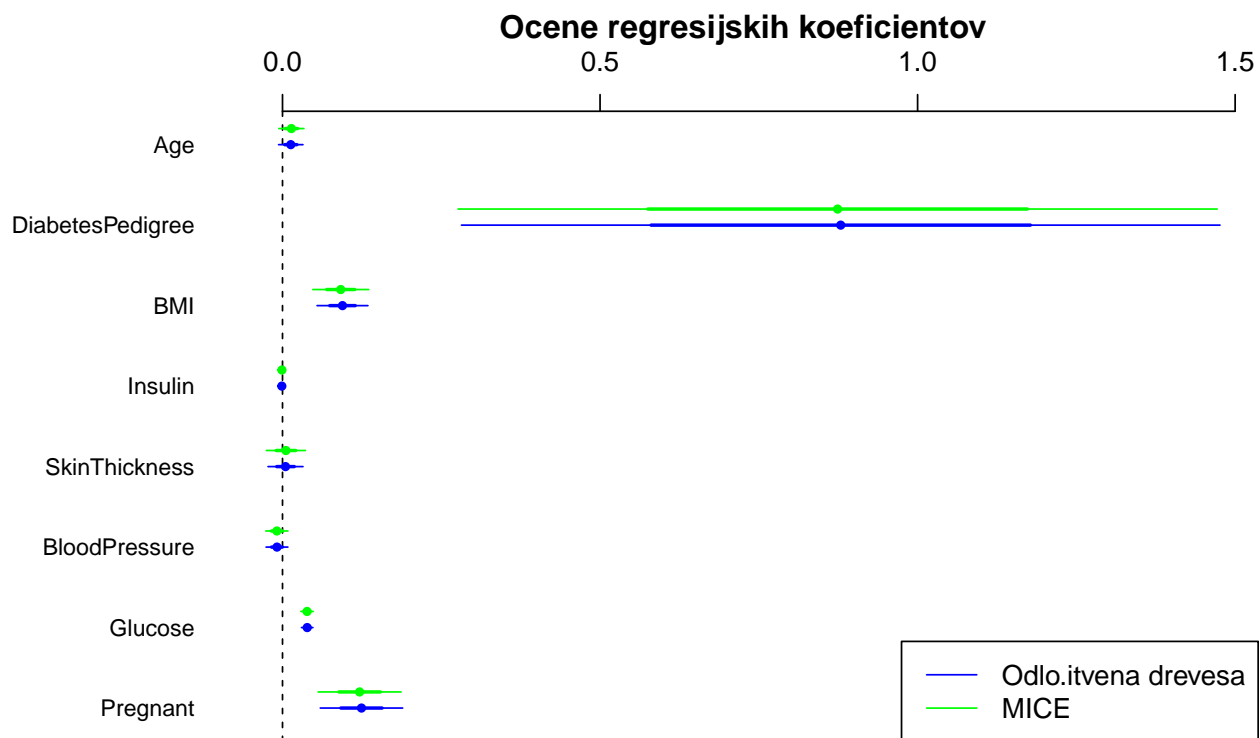
Poglejmo si sedaj katera metoda imputacije manjkajočih vrednosti se je odrezala najboljše glede na ocene regresijskih koeficientov in njihovih intervalov zaupanja.

Takoj opazimo, da sta original in listwise metodi enaki, kar se tiče intervalov zaupanja in ocen koeficientov, kar je logično, saj funkcija linearne regresije deluje po principu listwise metode.



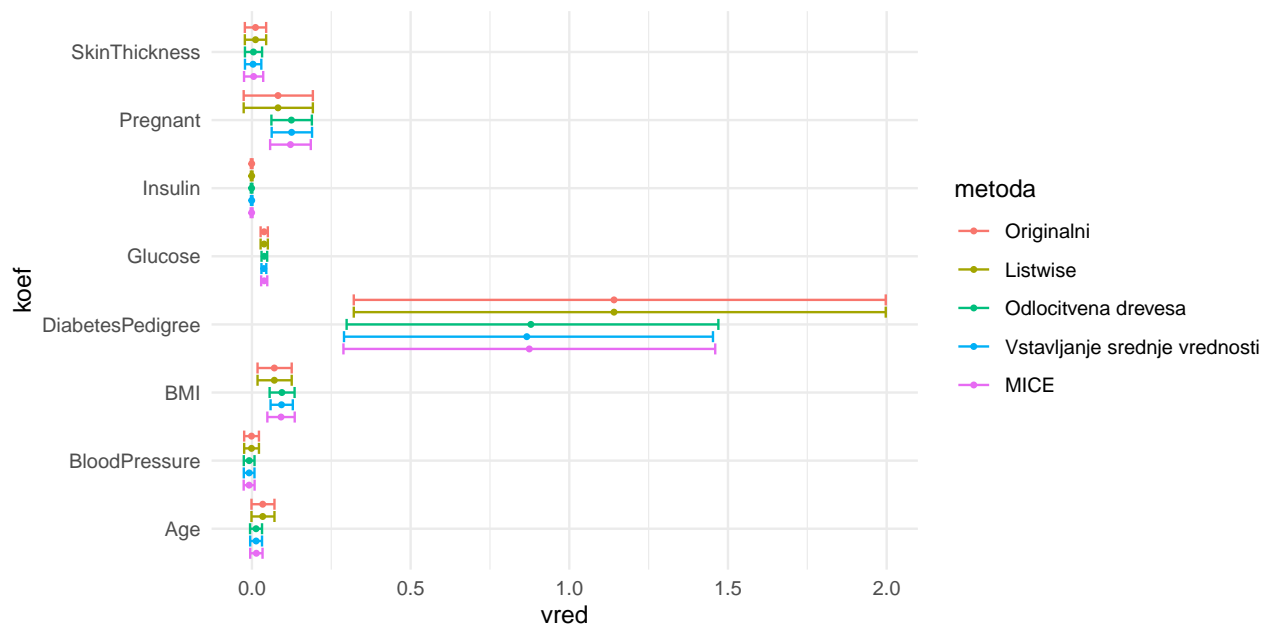
Slika 6: Prikaz ocen regresijskih koeficientov in intervalov zaupanja za vse metode imputacij manjkajočih vrednosti.

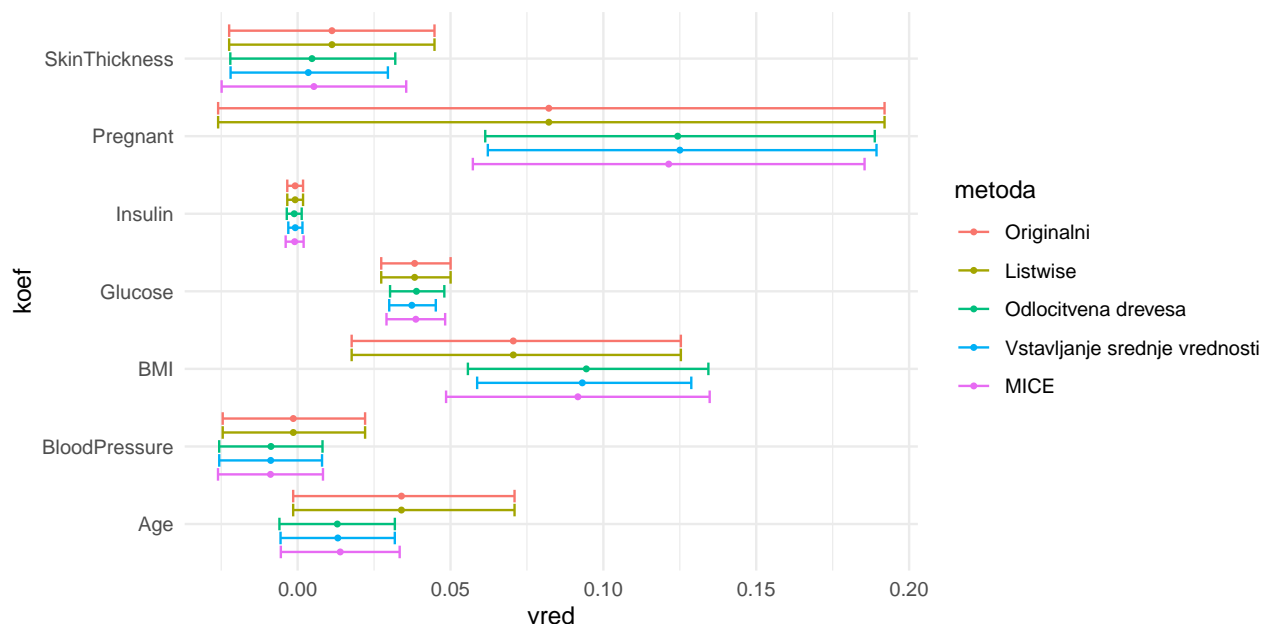
Podrobneje si pogledjmo metode imputacije manjkajočih vrednosti odločitvena drevesa in multiple(stohastične) imputacije(MICE) na spodnjem grafu. Vidno je, da sta si metodi zelo podobni oziroma ni večjih razlik med njima.



Slika 7: Prikaz ocen regresijskih koeficientov in intervalov zaupanja za dve metodi imputacij manjkajočih vrednosti.

Pri spremenljivkah, ki so imele največ manjkajočih vrednosti, npr *Insulin* ali *SkinThickness* ne vidimo večjih razlik v ocenah regresijskih koeficientov, tudi intervali zaupanja so dokaj ozki.





6 Zaključek

Glede na rezultate najne analize bi lahko metode razdelila v dve skupini. V prvi skupini je uporaba funkcije `glm()` na originalnih podatkih in uporaba *listwise* oziroma *pairwise* metode. Za našete primere sva pričakovala nekoliko slabše rezultate, saj pri oblikovanju logističnega modela uporabijo le podatke, ki so na voljo. S tem lahko pričakujemo drugačne vrednosti mer središčnosti in nekoliko višje mere razpršenosti, kar pa vpliva tudi na oceno koeficientov z logistično regresijo. To se je izkazalo res tudi v našem primeru, saj so se ocene koeficientov in 95% intervalov zaupanja znotraj skupine popolnoma ujemali. To pa tudi posledica delovanja funkcije `glm()`, saj pred modeliranje odstrani vse vrstice z manjkajočimi vrednostmi.

V drugo skupino pa bi lahko uvrstila vse preostale metode, ki sva jih še analizirala. To so *Odlocitvena drevesa*, *Vstavlanje srednje vrednosti* in *Multiple(stohastične) imputacije preko verižnih enačb*. Vse tri metode za obravnavo manjkajočih vrednosti so pri modeliranju dala podobne ocene koeficientov in njihovih 95% intervalov zaupanja. Mogoče bi lahko izpostavila nekoliko ožje intervale zaupanja pri metodi *Vstavlanje srednje vrednosti*, vendar je to tudi pričakovano, saj manjkajoče vrednosti nadomestimo s srednjo vrednosto spremenljivke. S tem podcenimo variabilnost, saj v primeru velikega števila manjkajočih vrednosti spremenljivke, bo večina vrednosti enakih, torej precej podcenimo variabilnost.

Po zgornji analizi bi za nadomestitev, manjkajočih vrednosti uporabila metodo *Odlocitvena drevesa* ali *Multiple(stohastične) imputacije preko verižnih enačb*.