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 mankajočih enot(več kot 30%), zato bova v nalogi definirala oz. ocenila, kakšen je mehanizem mankajočih vrednosti(po Rubinu) in obravnavala mankajoče vrednosti po treh različnih metodah.
Zanima naju torej mehanizem mankajočih vrednosti(zakaj in kako manjkajo podatki v podatkovnem nizu), da bova potem ustrezno izbrala metodo imputacije in jo primerjala še z drugima dvema metodama, ki se nama bosta zdeli primerni.

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` (debelo tkivo): Debelina kožnega gubca (v mm) na tricepsu, merjeno za testiranje telesne maščobe.
- `Insulin` (inzulin): Raven inzulina (v μU/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.

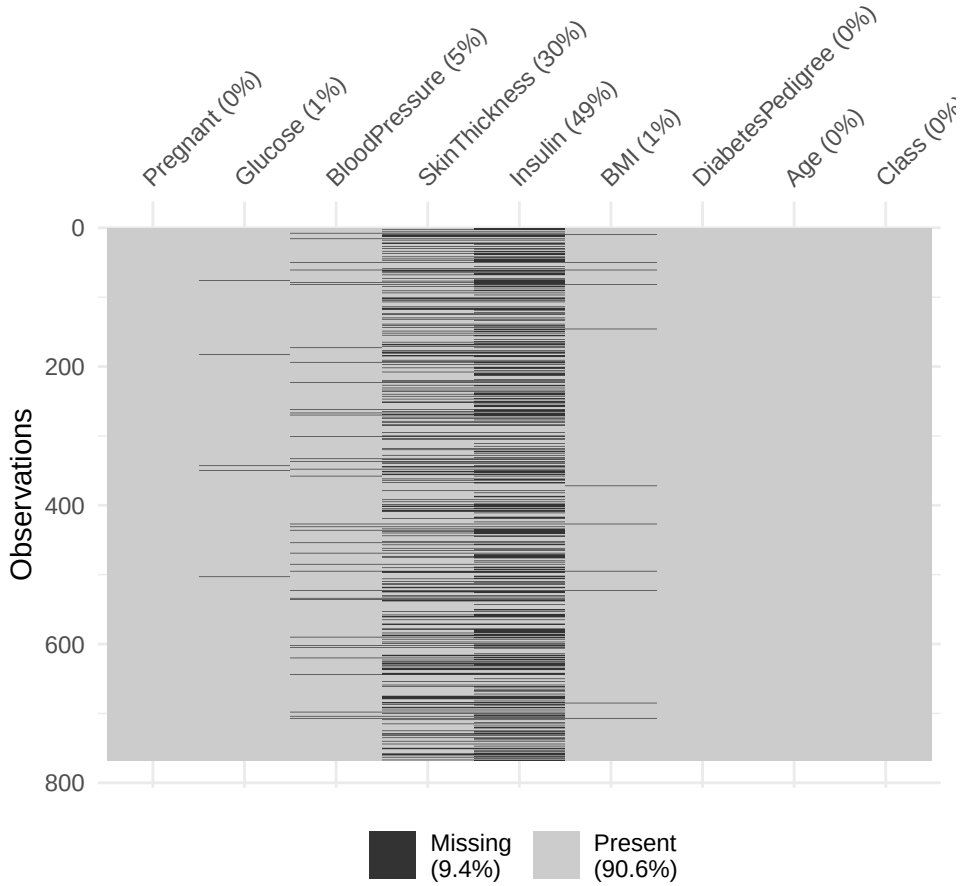
Poglejmo si osnovne statistike podatkov, iz katerih vidimo nekatere izjemne vrednosti, kot so npr. število nosečnosti 17, vrednost inzulina 846, debelost kože 99, ... Ker naju zanima analiza mankajočih vrednosti, podatkov podrobneje ne bova analizirala.

Tabela 1: Opisne statistike podatkov.

	n	mean	sd	median	min	max	se
Pregnant	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

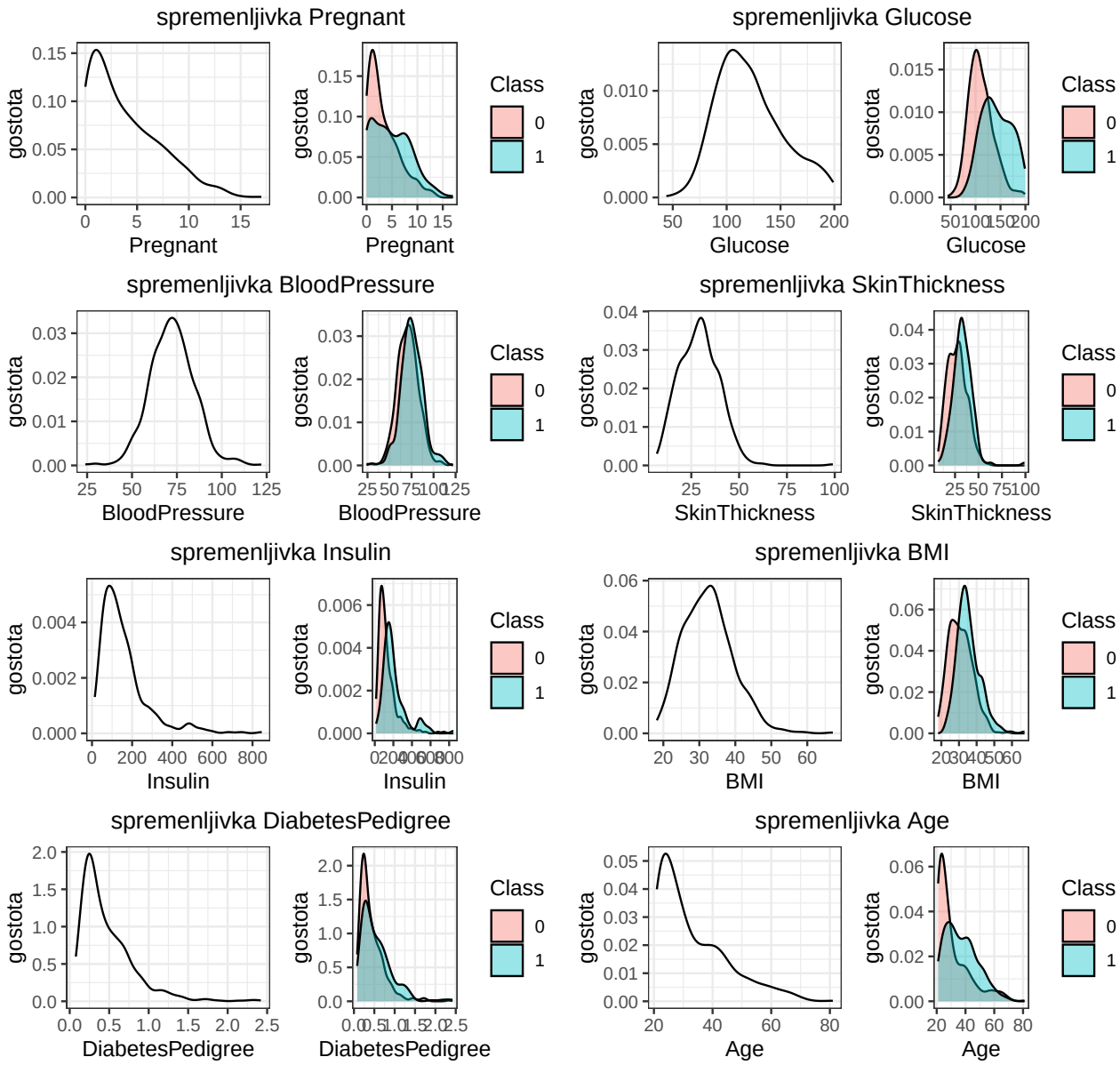
Class	768	0.35	0.48	0.00	0.00	1.00	0.02

Podatkovni niz vsebuje 768 primerov in torej vsak primer vključuje 8 merjenih spremenljivk in diagnozo(`Class`) - oseba ima diabetes(vrednost 1), oseba nima diabetesa(vrednost 0). Podatki vsebujejo veliko količino mankajočih vrednosti, kar lahko vidimo na spodnjem grafu glede na posamezno spremenljivko. Vidimo, da spremenljivke, kot so starost(`Age`), število nosečnosti(`Pregnant`), diagnoza(`Class`) in genetska nagnjenost(`DiabetesPedigree`) nimajo mankajočih vrednosti, torej imamo mankajoče vrednosti samo pri spremenljivkah, ki so bile verjetno izmerjene s strani medicinskega osebja.



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

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

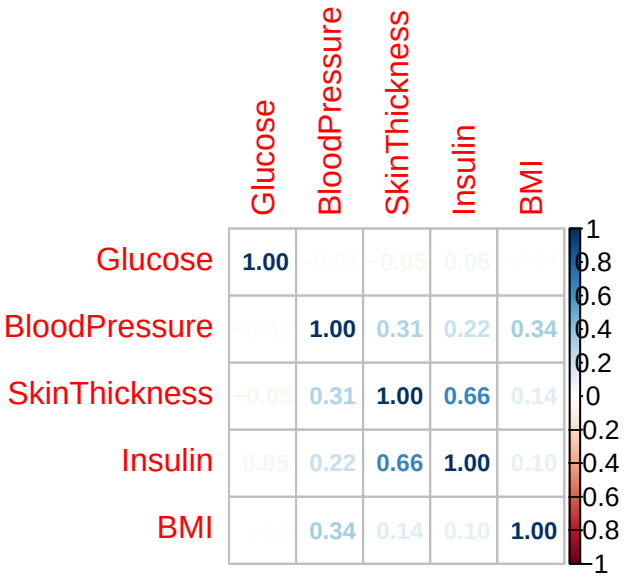


Slika 2: Porazdelitev spremenljivke v podatkovnem nizu, kjer je Class ciljna spremenljivka, ki označuje, ali ima posameznica diabetes (1) ali ne (0).

3 Mehanizem mankajočih vrednosti

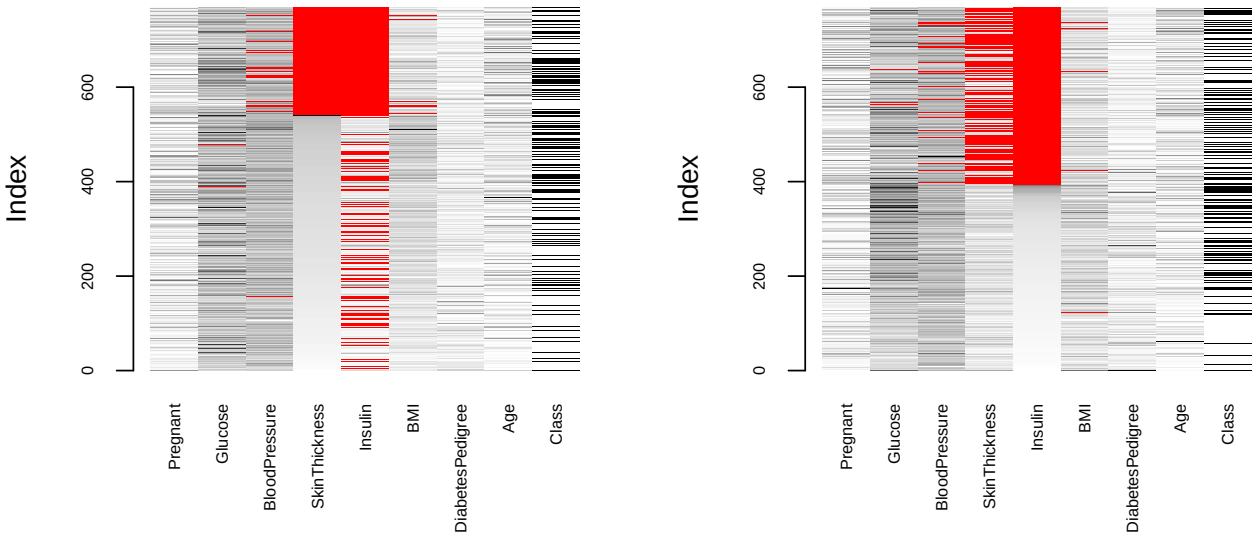
Za začetek si poglejmo korelacije med mankajočimi vrednostmi na spodnji korelacijski matriki. Opazimo nekaj pozitivnih korelacij, kar pomeni, da če imamo mankajočo vrednost pri eni spremenljivki, obstaja večja verjetnost, da bo tudi pri drugi.
Torej če imamo mankajočo vrednost pri spremenljivki `Insulin`, potem obstaja večja verjetnost, da bomo imeli mankajočo vrednost tudi pri spremenljivki `SkinThickness`(tudi pri `BloodPressure`). Če imamo mankajočo vrednost pri spremenljivki `SkinThickness`, obstaja večja verjetnost, da bomo imeli mankajočo vrednost še pri `BloodPressure`. Če pa bomo imeli mankajočo vrednost pr spremenljivki `BMI`, potem obstaja večja verjetnost, da bomo imeli mankajočo 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.



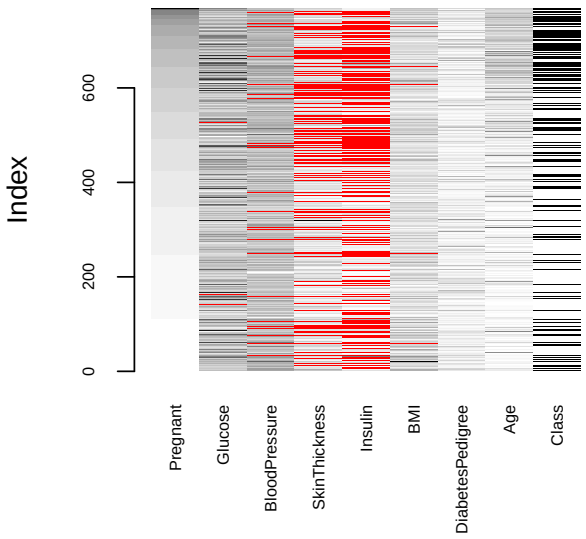
Slika 3: Korelacije med mankajočimi vrednostmi.

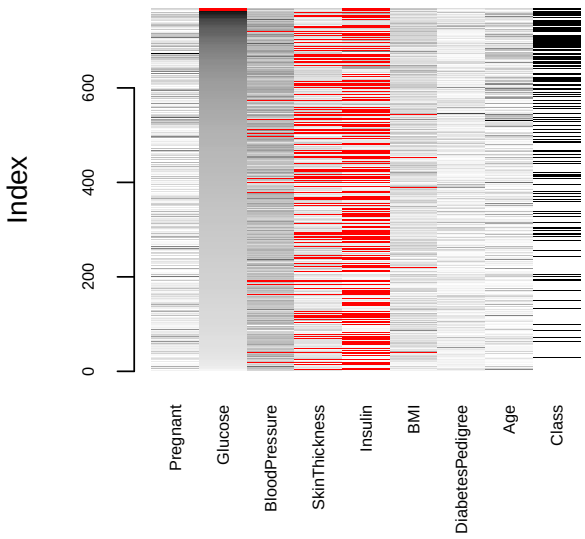
Poglejmo si še vzorce mankajočih vrednosti.
Najprej si poglejmo glede razvrstitev spremenljivk, kjer imamo največ mankajočih vrednosti, torej spremenljivki `SkinThickness` in `Insulin`. Iz vzorev na spodnjih grafih bi lahko rekli, da če imamo vrednost pri spremenljivki `Insulin`, jo imamo tudi pri drugih spremenljivkah(le dve izjemi), če pa nimamo vrednosti pri spremenljivi `SkinThickness`, jo nimamo tudi pri spremenljivki `Insulin` in obstaja velika verjetnost, da je ne bomo imeli tudi pri spremenljivki `BloodPressure`.



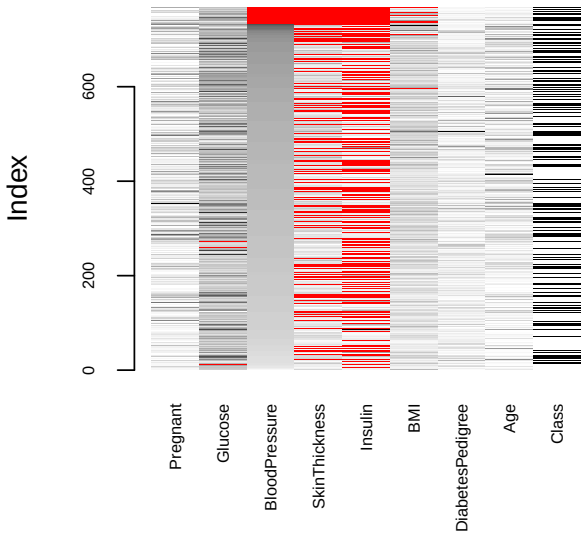
Ker je večina mankajočih vrednostih pri spremenljivkah `SkinThickness` in `Insulin`, nas sicer zanimajo predvsem razvrstitve glede na druge spremenljivke, zato so to tudi oglejmo.

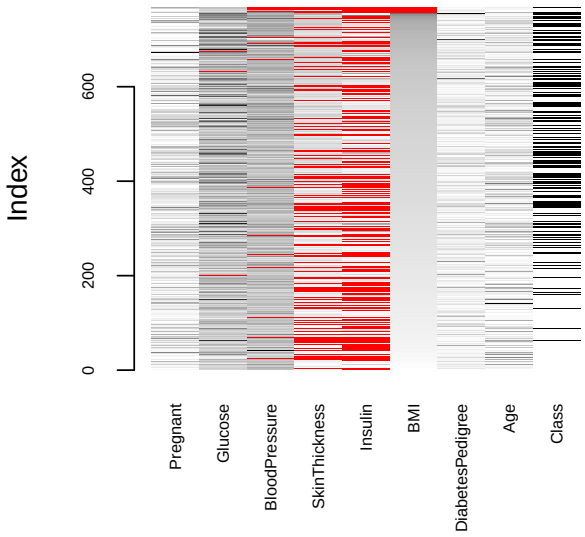
Iz spodnjega grafa imamo občutek, da obstaja večja verjetnost mankajočega podatka pri spremenljvkah `Insuline` in `SkinThickness` ob določenem številu nosečnostih pri ženskah. Pri razvrstitvi glede na spremenljivko `Glucose` ne zaznamo nekega vzorca.



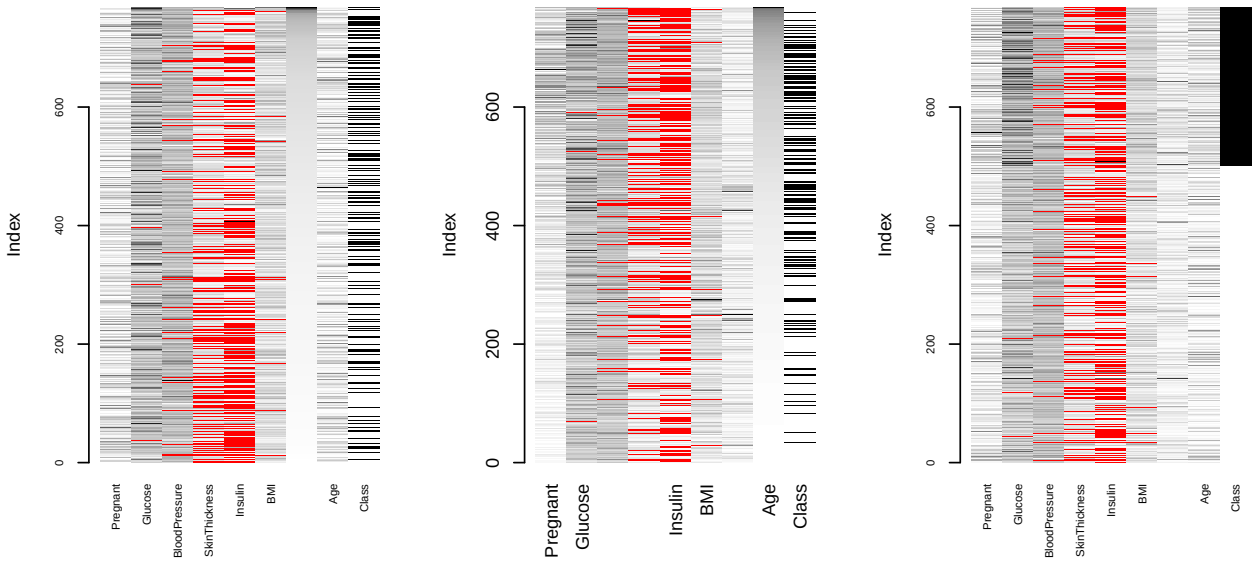


Iz spodnjega grafa se nam zdi, da če imamo mankajočo vrednost pri spremenljivki `BloodPressure`, potem obstaja velika verjetnost, da bomo imeli mankajoče vrednosti tudi pri spremenljivkah `SkinThickness` in `Insulin`. Prav tako obstaja večja verjetnost mankajoče vrednosti pri teh dveh spremenljivkah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





Na spodnjih treh grafih pa lahko vidimo, da je pri manjših vrednostih spremenljivke `DiabetesPedigree` in višji starosti več mankajočih 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 : Povprečji 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 mankajoči podatki pri spremenljivki `Insulin` niso naključni, saj rezultati testov nakazujejo, da so mankajoče vrednosti povezane s številom nosečnosti, da obstaja večja verjetnost mankajočega podatka pri starejših ženskah, pri tistih z višjim krvnim tlakom(`BloodPressure`) in nižjimi vrednostmi spremenljivke `DiabetesPedigree`. Mankajoči 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 mankajočih 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 mankajoči podatki `SkinThickness` povezani še z višjo povprečno starostjo(`Age`) in nižjimi vrednostmi `DiabetesPedigree`. Torej odsotnos 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 mankajočih vrednosti.
Pri mankajočih vrednostih spremenljivke smo testirali, ali na mankajoče 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 mankajočih 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 mankajočo 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 mankajoče vrednosti pri spremenljivkah niso statistično povezane z razredom `Class`. Razred, ki označujej prisotnost ali odsotnost sladkorne bolezni, torej ne vpliva na to, ali manjkajo podatki za spremenljivke.

4 Logistična regresija in imputacija mankajočih vrednosti

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 mankajočih vrednosti. Najprej bova preverila kakšne vrednosti koeficientov dobimo v primeru, da uporabimo podatke takše 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(stohastične) 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 mankajočih 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 mankajočih 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 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 številske spremenljivke in nekaj ekstremnih vrednosti, ki so osamelci(lahko vidimo v tabeli opisnih statistik podatkov na začetku). Odločila sva se za deterministični pristop vstavjanja mankajočih vrednosti, torej, da vstavljava le napovedi(brez slučajne napake modela).

Metoda deluje iterativno - pri vsakem koraku izboljša imputirane vrednosti na osnovi modela, ki ga gradi z uporabo predhodno imputiranih podatkov. Z večanjem števila iteracij običajno dobimo boljše rezultate.

Izvedla sva imputacije na podlagi 1 iteracije, petih in desetih. Ker najin nabor podatkov ni tako zelo ogromno velik, kot so ponavadi podatki pri strojnem učenju(kamor spadajo odločitvena drevesa), se nama zdi, da bo 10 iteracij več kot dovolj, pri dveh iteracijah pa se verjetno imputirane vrednosti še ne bodo stabilizirale.

V spodnjih tabelah lahko vidimo primerjavo med osnovnimi statistikami pri originalnih podatkih in podatkih z imputiranimi mankajočimi vrednostmi.

Tabela 3: Opisne statistike podatkov pri dveh iteracijah.

	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	121.61	30.46	1.10	763	121.69	30.54	1.11
BloodPressure	768	72.36	12.14	0.44	733	72.41	12.38	0.46
SkinThickness	768	28.88	9.36	0.34	541	29.15	10.48	0.45
Insulin	768	154.31	98.48	3.55	394	155.55	118.78	5.98
BMI	768	32.43	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

Tabela 4: Opisne statistike podatkov pri petih iteracijah.

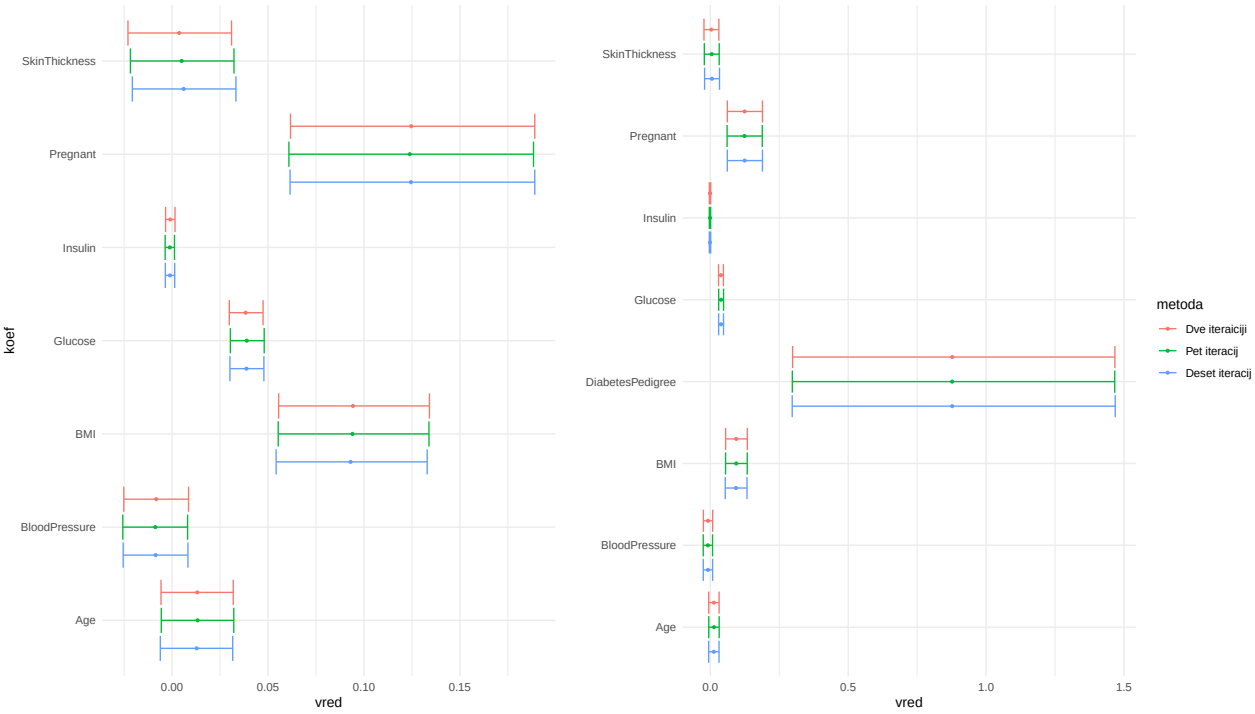
	odločitvena drevesa				izvirni podatki			
	n	mean	sd	se	n	mean	sd	se
Class*	768	1.35	0.48	0.02	768	0.35	0.48	0.02
Pregnant	768	3.85	3.37	0.12	768	3.85	3.37	0.12
Glucose	768	121.62	30.48	1.10	763	121.69	30.54	1.11
BloodPressure	768	72.35	12.15	0.44	733	72.41	12.38	0.46
SkinThickness	768	28.88	9.44	0.34	541	29.15	10.48	0.45
Insulin	768	154.13	98.84	3.57	394	155.55	118.78	5.98
BMI	768	32.41	6.90	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

Tabela 5: Opisne statistike podatkov pri desetih iteracijah.

	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	30.49	1.10	763	121.69	30.54	1.11
BloodPressure	768	72.36	12.15	0.44	733	72.41	12.38	0.46
SkinThickness	768	28.90	9.43	0.34	541	29.15	10.48	0.45
Insulin	768	154.38	99.91	3.61	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

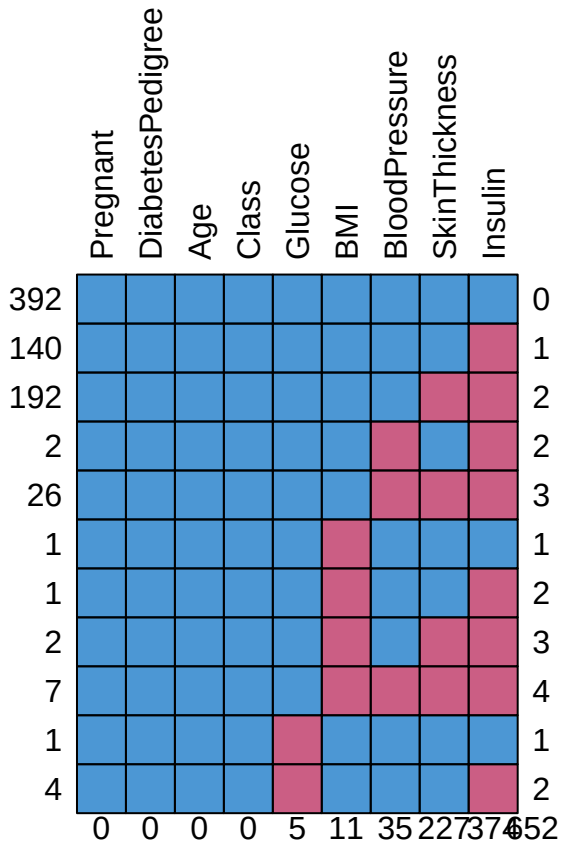
Tokrat lahko vidimo, da smo z izračunom nadomestili manjkajoče vrednosti (število vrednosti v vsakem stolpcu (spremenljivki) je enako 768, tako kot na izvirnih podatkih) pri vseh iteracijah seveda. Pričakovano so se spremenile vrednosti opisnih spremenljivk, predvsem pa opazimo razliko oz. zmanjšanje 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 mankajočih vrednosti(npr. `Insulin`), kar je seveda logično, saj moramo vstaviti več mankajočih vrednosti. Večjih razlik med številom iteracij pa ni, kar lahko vidimo tudi na spodnjih dveh grafih, zato je morda privzeta vrednost(deset iteracij) najboljša izbira.



Slika 4: Prikaz ocen regresijskih koeficientov in intervalov zaupanja za metodo odločitvenih dreves imputacij mankajočih vrednosti za različne iteracije(brez spremenljivke DiabetesPedigree levo in z desno).

4.3 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.

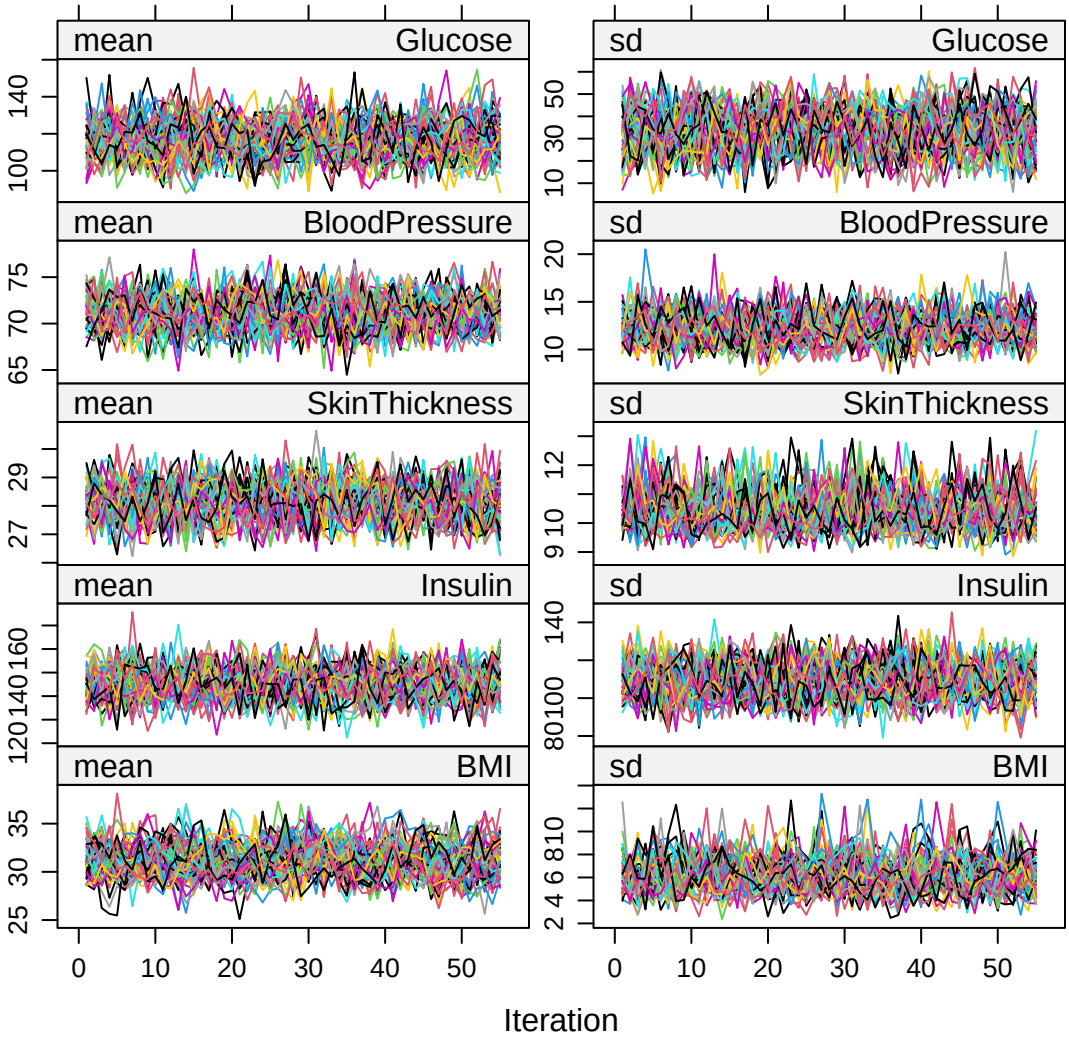


Slika 5: Tabela mankajočih vrednosti.

Enotam z manjkajočimi vrednostmi najpogosteje manjkajo vrednosti spremenljivk `Insulin`(1), `SkinThickness`(2) ali `BloodPressure`(3), kar smo ugotovili že pri mehanizmu mankajočih 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% mankajočih enot) in `maxit` = 55 (nekoliko višja vrednost, da vrednosti zagotovo skonvergirajo).

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



Slika 6: Grafi povprečja in standardnega odklona spremenljivk z mankajočimi 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 mankajočih vrednosti se je odrezala najbolje glede na ocene regresijskih koeficientov in njihovih intervalov zaupanja. Na enem grafu bomo prikazali spremenljivko `DiabetesPedigree`, na drugem pa ostale spremenljivke zaradi boljšega pregleda.

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. Opazno je tudi, da je širina intervalov pri vseh spremenljivkah ožja pri metodi imputacije z odločitvenimi drevesi in MICE. Vse tri metode so si sicer dokaj podobne, še posebaj pri spremenljivki Insulin, kjer smo imeli največ mankajočih vrednosti, pri spremenljivki SkinThickness je vseeno razlika med metodo listwise in odločitvenimi drevesi ali metodo MICE.
Pri spremenljivkah, ki so imele največ mankajočih vrednosti, npr `Insulin` ali `SkinThickness` torej ne vidimo večjih razlik v ocenah regresijskih koeficientov, tudi intervali zaupanja so dokaj ozki glede na metode imputacije mankajočih vrednosti. S tem ko pri spremenljivkah, ki pa niso imele mankajoče vrednosti, pa

dobimo ožje intervale zaupanja če v linearni regresiji uporabimo podatke, kjer smo imputirali mankajoče vrednosti pri drugih spremenljivkah s pomočjo metode odločitvenih dreves in MICE. Torej metodi pozitivno prispevata k ocenjevanju koeficientov pri linearni regresiji in njihovim intervalom zaupanja.

Slika 7: Prikaz ocen regresijskih koeficientov in intervalov zaupanja za vse metode imputacij mankajočih vrednosti(brez spremenljivke DiabetesPedigree levo in z desno).

6 Vstavljanje srednje vrednosti

Med tem, da odstranimo vse enote z mankajočimi vrednostmi v katerem koli stolpcu in tem, da jih imputiramo s pomočjo dveh ustreznih metod sicer obstaja razlika. Ampak zgolj iz radovednosti pa naju zanima ali sta metodi odločitvenih dreves in MICE res dobri oziroma boljši od npr. vstavljanja srednje vrednosti(povprečja), ki sicer ni primerna metoda praktično nikoli, saj podcenjuje variabilnost in korelacijo med podatki. Je pa sicer metoda, ki je dokaj preprosta, ni računsko zahtevna in manjkajoče vrednosti nadomestimo na dokaj lahek način.

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 spremneljivki, pričakujeva da bo delovala najslabše, vendar naju vseeno zanima koliko se bo zares razlikovala od preostalih metod.

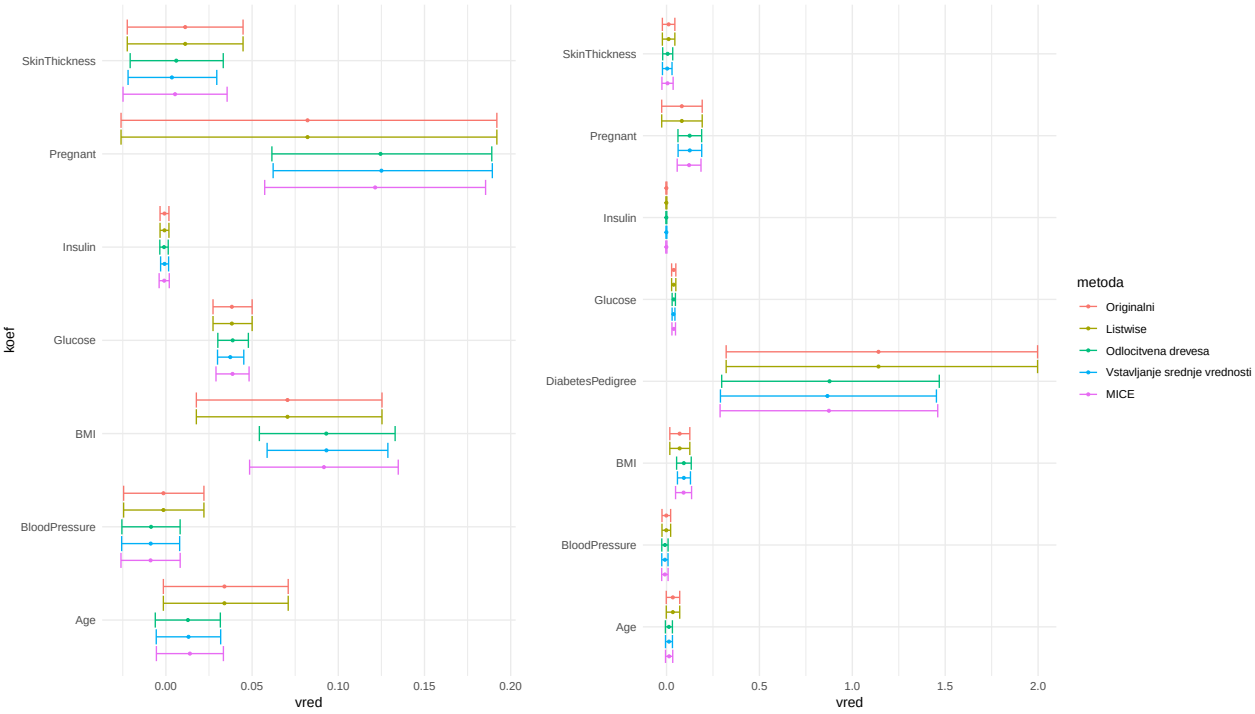
Tabela 6: Opisne statistike podatkov.

	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 veliko spremembo standardnega odklona pri spremenljivkah z veliko manjkajočimi vrednostmi kot sta `SkinThickness`, `Insulin`, pri ostalih pa sprememba ni tako zelo velika.

Na spodnjem grafu(ponovno je prikaz ločen najprej za izbrane spremenljivke in potem še za vse, da je bolj pregledno) pa se na prvi pogled zdi, da metoda ni tako zelo slaba, ker ima dokaj ozke intervale zaupanja(tudi ožje kot ostale metode), ampak to pričakujemo, predvsem pri spremenljivkah z veliko mankajočimi vrednostmi, ker bo potem veliko enot imelo enako vrednost in s tem odstranimo variabilnost; enako sicer lahko opazimo tudi pri spremenljivkah z malo mankajočimi vrednostmi. Torej to, kljub ozkim intervalom zaupanja, ni najboljša metoda imputacije mankajočih vrednosti.



Slika 8: Prikaz ocen regresijskih koeficientov in intervalov zaupanja za vse metode imputacij mankajočih vrednosti(brez spremenljivke DiabetesPedigree levo in z desno).

7 Zaključek

Glede na rezultate najne analize bi lahko metode razdelila v dve skupini. V prvi skupini je uporaba funckije `glm()` na originalnih podatkih in uporaba *listwise* oziroma *pairwise* metode. Za naštete 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 najnem 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, torej sta to torej enaki metodi.
V drugo skupino pa bi lahko uvrstila vse preostale metode, ki sva jih še analizirala. To so *Odlocitvena drevesa*, *Vstavljanje 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(predvsem odločitvena drevesa in MICE). Res imamo ožje intervale zaupanji pri metodi vstavljanja srednje vrednosti, vendar se to ne izkaže za dobro, saj pri spremenljivkah z veliko mankajočimi vrednostmi vstavimo veliko enakih vrednosti, kar zmanjša variabilnost oz. jo s tem podcenimo.

Po zgornji analizi bi, na podlagi grafov in rezultatov, za nadomestitev manjkajočih vrednosti uporabila metodo *Odlocitvena drevesa* ali *Multiple(stohastične) imputacije preko verižnih enačb*. Obe metodi se nama zdita primerni za najine podatke.