Univerzitet u Nišu
Elektronski fakultet
Katedra za računarstvo

Seminarski rad

NEDOSTAJUĆI PODACI

Student:
Petra Milosavljević, 1582

Profesor:
Doc. dr Aleksandar Stanimirović

Sadržaj

1. Uvod

U savremenim aplikacijama mašinskog učenja, kvalitet podataka predstavlja ključni faktor za uspeh modela i tačnost predikcija. Ipak, u realnim situacijama, skupovi podataka su često nepotpuni iz različitih razloga. To mogu biti tehnički problemi prilikom prikupljanja, ljudske greške ili namerni propusti.

Kako većina statističkih modela funkcioniše samo na potpunim opservacijama, neophodno je rešiti problem nedostajućih podataka, bilo brisanjem nepotpunih opservacija ili zamenom nedostajućih vrednosti procenjenom vrednošću na osnovu dostupnih informacija, procesom koji se zove imputacija. Obe metode mogu značajno uticati na zaključke koje možemo izvući iz podataka. Nedostajući podaci nisu samo tehnički problem već i ozbiljan izazov u domenu analize podataka, jer mogu dovesti do pristrasnosti, smanjenja statističke snage i netačnih zaključaka. Zbog toga je važno razumeti prirodu nedostajućih podataka, kako bi se odredila metoda koja se koristi za njihovo rešavanje.

Razumevanje šta su nedostajući podaci, kako nastaju i zašto je važno pravilno ih obraditi je od ključnog značaja kada se radi sa podacima iz stvarnog sveta, posebno sa tabelarnim podacima, koji su jedan od najčešće korišćenih tipova podataka u stvarnom svetu. U literaturi su definisana tri mehanizma nedostajućih podataka: Missing Completely At Random (MCAR), Missing At Random (MAR) i Missing Not At Random (MNAR), pri čemu svaki predstavlja jedinstvene izazove za imputaciju. Većina postojećih radova fokusira se na MCAR, koji je relativno lak za obradu. Specifični mehanizmi nedostajućih podataka, kao što su MNAR i MAR, manje su istraženi i shvaćeni. U nekim slučajevima MCAR mogu biti bezbedno ignorisani, dok MAR i MNAR podaci zahtevaju pažljiviju analizu i naprednije metode kako bi se smanjila pristrasnost i očuvala tačnost modela.

Cilj ovog rada je da pruži pregled glavnih koncepata i tehnika koje se odnose na upravljanje nedostajućim podacima, kao i da se istraži njihov uticaj na modele mašinskog učenja. Posebna pažnja će biti posvećena metodama za imputaciju podataka. Ovaj rad će predstaviti relevantne metode koje se koriste u procesu prikupljanja, obrade i modeliranja podataka sa nedostajućim vrednostima, uz analizu njihovih prednosti i nedostataka.

2. Tipovi nedostajućih podataka

Načini na koje podaci nedostaju utiču na određene pretpostavke koje podržavaju naše metode imputacije podataka. Tri glavna tipa nedostajućih podataka mogu se opisati u zavisnosti od odnosa između posmatranih (dostupnih) i neposmatranih (nedostajućih) podataka. Radi jednostavnosti, razmotrićemo slučaj nedostajućih podataka u jednoj varijabli. Da bismo matematički definisali nedostajuće podatke, dataset X može se podeliti na dva dela:

$$X = \{X_o, X_m\}$$

gde X_o odgovara posmatranim podacima, a X_m nedostajućim podacima u datasetu. Za svaku opservaciju definišemo binarni odgovor koji ukazuje na to da li ta opservacija nedostaje ili ne:

$$R = \begin{cases} 1 & if \quad X\,observed \\ 0 & if \quad X\,missing \end{cases}$$

Tip nedostajuće vrednosti može se razumeti kroz verovatnoću da opservacija nedostaje, $Pr(R)$, s obzirom na posmatrane i nedostajuće opservacije, u obliku:

$$Pr(R|x_o, x_m)$$

Tri tipa nedostajućih podataka zavise od toga da li verovatnoća odgovora R zavisi ili ne zavisi od posmatranih i/ili nedostajućih vrednosti i dele se na:

1. **Nedostajući podaci potpuno nasumično (MCAR)**

U ovom slučaju, nedostajući podaci se pojavljuju potpuno nasumično, bez ikakve veze sa vrednostima drugih varijabli ili sa vrednostima koje nedostaju. Verovatnoća da podatak bude nedostajući zavisi samo od samog podatka i svodi se na $Pr(R|x_o, x_m) = Pr(R)$.

Na primer, zamislite da doktor zaboravi da zabeleži pol svakog šestog pacijenta koji uđe u ordinaciju. Ne postoji skriveni mehanizam povezan sa bilo kojom varijablom i ne zavisi od bilo koje karakteristike pacijenata.

Ako podaci nedostaju na ovaj način, oni se mogu jednostavno zanemariti bez većih posledica po analizu, jer nisu sistematski povezani ni sa jednom varijablom. Algoritmi mašinskog učenja koji pretpostavljaju da podaci nedostaju nasumično neće biti pristrasni, ali se može izgubiti statistička snaga jer se smanjuje broj dostupnih uzoraka.

MCAR podaci se mogu obrisati ili nadomestiti jednostavnim metodama imputacije (npr. popunjavanje prosekom, medianom), jer ovo neće značajno uticati na rad modela.

2. **Nedostajući podaci nasumično (MAR)**

U ovom slučaju, verovatnoća da neka vrednost nedostaje zavisi od drugih posmatranih podataka, ali ne zavisi od same vrednosti koja nedostaje. Drugim rečima, postoji neka veza između nedostajućih vrednosti i drugih dostupnih varijabli, ali ne i sa nedostajućom vrednošću. Posmatrani podaci su statistički povezani sa nedostajućim varijablama i moguće je proceniti nedostajuće vrednosti na osnovu posmatranih podataka. Ovaj slučaj nije potpuno 'nasumičan', ali je najgeneralniji slučaj gde možemo zanemariti mehanizam nedostajanja, jer kontrolišemo informacije na kojima zavisi nedostajanje, odnosno posmatrane podatke. Drugim rečima, verovatnoća da neki podatak nedostaje za određenu varijablu ne zavisi od vrednosti te varijable, nakon prilagođavanja za posmatrane vrednosti. Matematički, verovatnoća nedostajanja svodi se na $Pr(R|x_o, x_m) = Pr(R|x_o)$.

Zamislite da ako stariji ljudi manje verovatno obaveste doktora da su imali pneumoniju, stopa odgovora na varijablu "pneumonija" će zavisiti od varijable starosti. Odnosno, ovde se razlog zašto podaci nedostaju može objasniti faktorom, tj. godinama pacijenta.

Podaci koji nedostaju uslovno na druge varijable mogu izazvati pristrasnost ako se ne tretiraju pravilno. Međutim, ako imate dovoljno informacija iz drugih dostupnih varijabli, moguće je primeniti imputaciju kako bi se nadomestile nedostajuće vrednosti.

MAR podaci se često mogu adekvatno tretirati metodama imputacije koje uzimaju u obzir vrednosti drugih varijabli. Na primer, regresione imputacione metode ili imputacija zasnovana na algoritmu kao što je k-NN (k najbližih suseda) mogu se koristiti kako bi se nadomestile vrednosti.

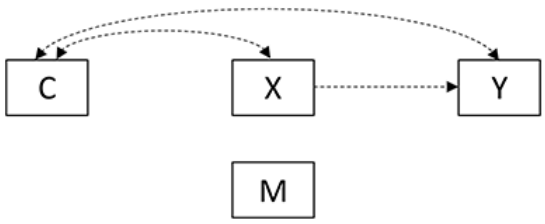
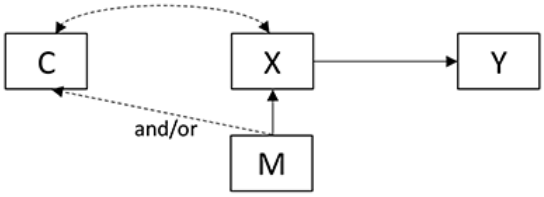
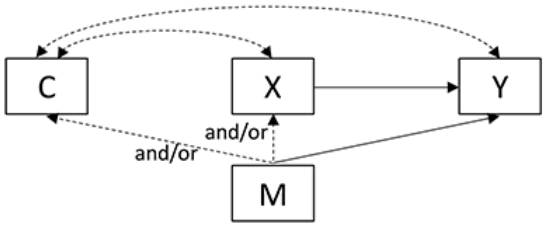
3. **Nedostajući podaci nisu nasumični (MNAR)**

Ovo se odnosi na slučaj kada nijedna od prethodnih pretpostavki (MCAR ili MAR) ne važi, već podaci nedostaju već iz specifičnih razloga koji su povezani sa vrednostima koje nedostaju. U ovom slučaju nedostajući podaci zavise kako od nedostajućih, tako i od posmatranih vrednosti. Određivanje mehanizma nedostajanja je obično nemoguće, jer zavisi od neviđenih podataka. Iz toga proizilazi važnost izvođenja analiza osetljivosti i testiranja šta se može zaključiti pod različitim pretpostavkama.

Na primer, možemo zamisliti da će pacijenti sa niskim krvnim pritiskom verovatnije ređe meriti krvni pritisak. Tako da će nedostajući podaci za varijablu "krvni pritisak" delimično zavisiti od vrednosti krvnog pritiska.

Možemo zaključiti da postoji neka sistemska pristrasnost koja može biti vrlo problematična za analizu. Nedostajuće podatke ovog tipa nije lako otkriti i to može značajno izmeniti rezultate. Ako se MNAR podaci ignorišu ili pogrešno imputiraju, to može dovesti do ozbiljnih pogrešnih zaključaka.

MNAR podaci su najteži za rešavanje, jer zahtevaju dodatne pretpostavke ili informacije iz spoljašnjih izvora. Jedan od pristupa je model-based imputacija, gde se razvijaju modeli koji pokušavaju da predvide nedostajuće vrednosti na osnovu drugih podataka i pretpostavki o prirodi nedostajanja.

Categorization	Model	Ignorable?	Statistical Methods for Addressing
Missing Completely at Random (MCAR)		Ignorable	None needed, except to preserve sample size
Missing at Random (MAR)		Ignorable	• Appropriate controls • Maximum likelihood • Multiple imputation • Bayesian techniques
Missing Not at Random (MNAR)		• Non-ignorable (generally) • Ignorable with availability of salient auxiliary variables	• Multiple imputation • Maximum likelihood with auxiliary variables • Bayesian techniques

Slika 1. Tipovi nedostajućih podataka

Sa šeme na slici 1. možemo i vizuelno videti zavisnosti za svaki od tipova nedostajućih podataka. Nedostajući podaci su označeni sa *M*, sa *X* posmatrani (observed) podaci, a sa *Y* neposmatrani (unobserved) podaci.

3. Obrada nedostajućih podataka

Metode treba prilagoditi datasetu od interesa, razlozima za nedostajuće vrednosti i proporciji nedostajućih podataka. Generalno, metoda se bira zbog svoje jednostavnosti i sposobnosti da uvede što manje pristrasnosti u dataset.

Kada su podaci MCAR ili MAR, možemo ignorisati razloge za nedostajuće podatke, što pojednostavljuje izbor metoda koje će se primeniti. U tom slučaju, bilo koja metoda može biti primenjena. Ipak, teško je zaključiti da li su podaci MCAR ili MAR. Validna strategija je ispitivanje osetljivosti rezultata na MCAR i MAR pretpostavke kroz poređenje nekoliko analiza, pri čemu razlike u rezultatima između analiza mogu pružiti određene informacije o tome koje pretpostavke su najrelevantnije.

Najčešćešćene metode za obradu se mogu pobediliti u tri osnovne grupe:

- Izbacivanje primera koji imaju nedostajuće vrednosti u svojim atributima. U ovu kategoriju spada i brisanje atributa sa visokim nivoima nedostajućih vrednosti. Ovaj pristup je jednostavan za primenu i pogodan kada je procenat nedostajućih podataka mali, ali može dovesti do gubitka značajnog dela podataka, što može negativno uticati na kvalitet analize.
- Korišćenje metode maksimalne verovatnoće, gde se prvo procenjuju parametri modela na osnovu potpunog dela podataka, a zatim se ti parametri koriste za imputaciju vrednosti putem uzorkovanja (*sampling*). Ovaj pristup omogućava precizniju imputaciju, posebno kod velikih datasetova, i čuva podatke, ali je kompleksan za implementaciju i može zahtevati puno resursa.
- Imputacija nedostajućih vrednosti, gde se nedostajuće vrednosti popunjavaju procenjenim. U većini slučajeva atributi podataka nisu nezavisni jedni od drugih, pa se identifikovanjem odnosa među njima mogu odrediti nedostajuće vrednosti. U ovom slučaju, zadržavaju se sve opservacije u datasetu i mogu se dobiti kvalitetne procene kada se odnosi među atributima pravilno identifikuju. Međutim, može uvesti pristrasnost i smanjiti varijabilnost ako imputacija nije dovoljno precizna.

3.1. Brisanje podataka (Deletion Methods)

Najjednostavniji način za rešavanje problema nedostajućih podataka je odbacivanje podataka ili opservacija koje imaju nedostajuće vrednosti. Generalno, metode brisanja vode do validnih zaključaka samo u slučaju MCAR podataka. Postoje dva osnovna načina za brisanje podataka: brisanje cele opservacije (Listwise Deletion) i brisanje parova podataka (Pairwise Deletion).

3.1.1. Listwise deletion (brisanje cele opservacije)

U ovom slučaju, sve opservacije sa barem jednom nedostajućom varijablom se odbacuju (Slika 2). Glavna pretpostavka je da će preostali podaci biti dobra reprezentacija populacije i da time model neće imati pristrasnost ka nekoj podgrupi. Ova pretpostavka je prilično restriktivna i podrazumeva MCAR mehanizam.

Gender	GLUCOSE	Age
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Slika 2. Primer listwise brisanja

Najveća prednost ove metode je njena jednostavnost, i uvek ima smisla koristiti je ukoliko je broj odbačenih opservacija relativno mali u poređenju sa ukupnim brojem. Glavni nedostaci su smanjena statistička snaga (zbog smanjenja broja uzoraka, procene će imati veće standardne greške), gubitak informacija i moguća pristrasnost analize, posebno ako podaci nisu MCAR.

3.1.2. Pairwise deletion (brisanje parova podataka)

Kod ove metode, opservacije se ne brišu u potpunosti. Umesto toga, u svakoj analizi se koriste samo one opservacije koje imaju sve potrebne vrednosti za varijable uključene u tu specifičnu analizu. Na primer, ako prva opservacija ima nedostajuću vrednost u j-toj koloni, ta opservacija će biti isključena samo kada se analizira j-ta karakteristika. Međutim, ona će biti uključena u analize koje se odnose na druge karakteristike.

Case Study		
S1	S2	S3
Gender	GLUCOSE	Age
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Slika 3. Primer pairwise brisanja

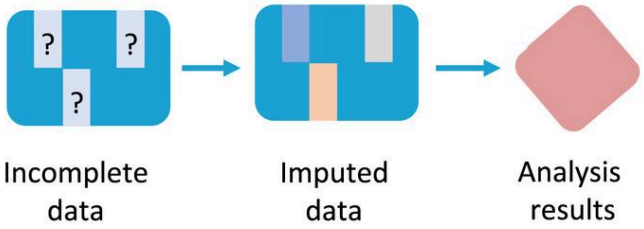
Glavne prednosti ovog pristupa u odnosu na listwise brisanje su to što očuva više podataka i bolje iskorišćava dostupne informacije. Međutim, kako se različiti uzorci koriste u različitim analizama, može dovesti do nekonzistentnih rezultata.

3.2. Imputacija podataka (Data Imputation)

Kako bi se prevazišla ograničenja metoda brisanja, metode imputacije igraju ključnu ulogu u rešavanju problema nedostajućih podataka. Tehnike imputacije imaju za cilj da povrate nedostajuće vrednosti, a da pritom očuvaju integritet celokupnog dataseta. Ove metode su posebno dragocene kada su dostupni uzorci podataka ograničeni ili kada se radi o podacima sa specifičnim mehanizmima nedostajanja. Metode imputacije uključuju popunjavanje nedostajućih vrednosti koristeći različite strategije, poput imputacije prosečne vrednosti, regresione imputacije i tehnika mašinskog učenja. Korišćenjem informacija iz posmatranih podataka, imputacija omogućava sveobuhvatnu analizu i smanjuje potencijalne pristrasnosti koje se javljaju prilikom uklanjanja podataka.

3.2.1. Jednostruka imputacija

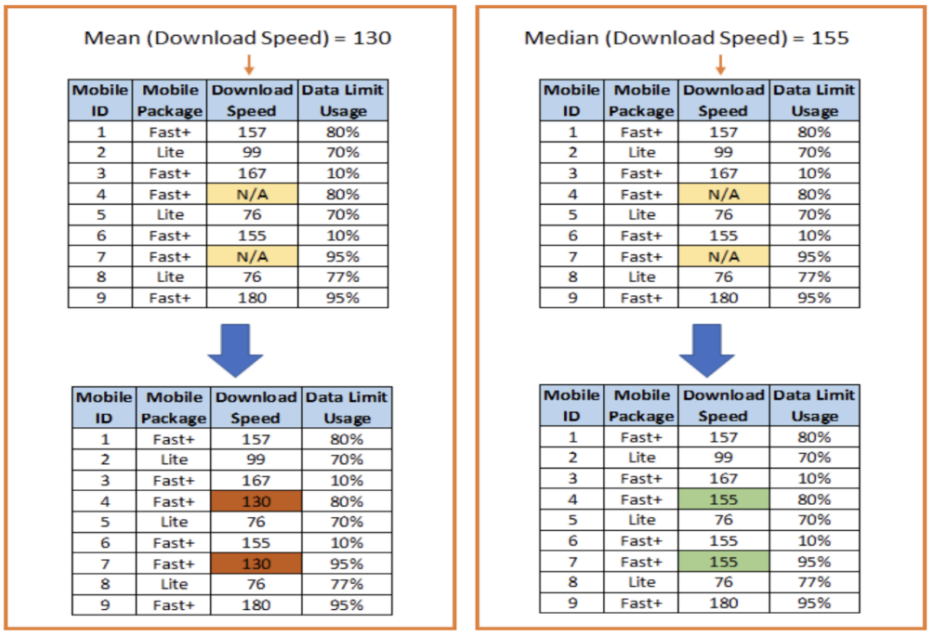
Jednostruka imputacija podrazumeva zamenu nedostajućih vrednosti jednom procenjenom vrednošću. U ovom pristupu, svaka nedostajuća vrednost se imputira jednom vrednošću, koristeći specifične pretpostavke ili statističke metode. Ovaj proces osigurava da svaka nedostajuća vrednost bude zamenjena jednom imputiranom vrednošću.



Slika 4. Jednostruka imputacija

Mean, Median i Mode

Metode imputacije koristeći srednju vrednost, medijanu i modus uključuju izračunavanje srednje vrednosti, medijane ili modusa nenedostajućih vrednosti unutar svake kolone, a zatim korišćenje tih vrednosti za imputaciju nedostajućih podataka. Srednja vrednost i medijana su pogodne za numeričke podatke, dok je modus pogodniji za kategorijske ili binarne podatke, jer nemaju srednju vrednost ili medijanu. Ove metode se mogu primeniti ne samo na tabelarne podattke, već i na druge formate, poput podataka o slikama i kod vremenskih serija, jer se mogu predstaviti u numeričkim formatima. Međutim, važno je napomenuti da, budući da se ove metode oslanjaju na jednostavne proračune, možda neće uhvatiti složenu strukturu raspodele podataka.



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%

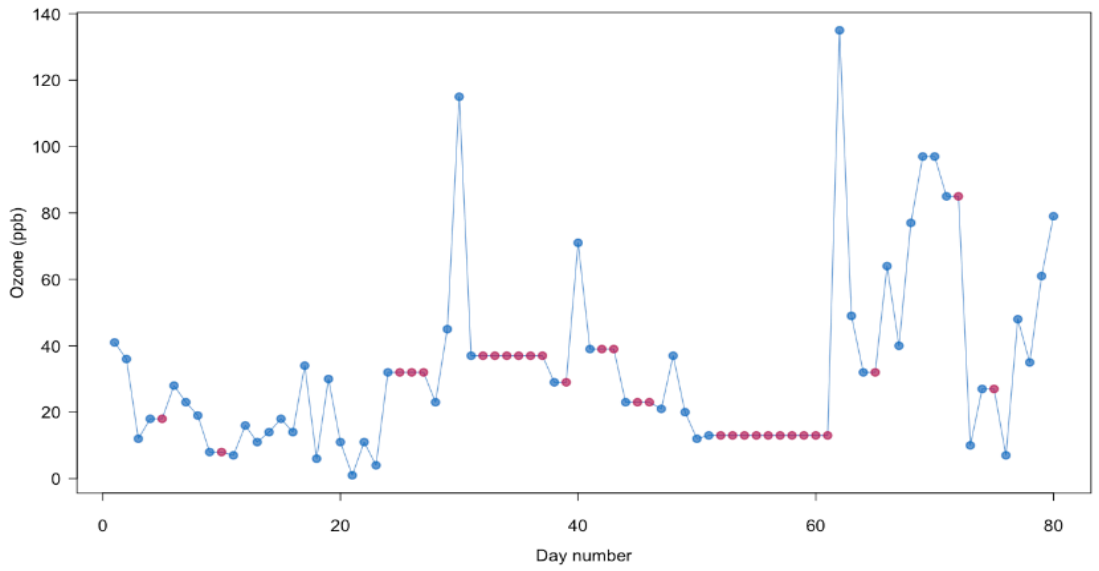
Slika 5. Primeri Mean i Median imputacije

Iako su metode imputacije koristeći srednju vrednost, medijanu i modus jednostavne za implementaciju, imaju određena ograničenja i efikasne su samo ako mehanizam nedostajanja podataka prati MCAR pretpostavku. Imputacija srednje vrednosti zanemaruje varijabilnost

podataka, jer svi nedostajući podaci dobijaju istu vrednost. Uprkos ovim nedostacima, u određenim datasetima i scenarijima, imputacija srednje vrednosti može nadmašiti druge tehnike imputacije.

LOCF i NOCB

Metode Last Observation Carried Forward (LOCF) i Next Observation Carried Backward (NOCB) koriste najbližu posmatranu vrednost pre ili posle nedostajuće vrednosti kako bi imputirale nedostajuće vrednosti. Ove metode se često primenjuju u vremenskim serijama podataka, gde se javljaju obrasci nedostajanja, a promenljive se ponovo mere kroz seriju vremenskih tačaka ili su posmatranja zavisna od najbližih tačaka. LOCF imputira nedostajuću vrednost vrednošću poslednjeg posmatranja, dok NOCB imputira nedostajuću vrednost vrednošću sledećeg posmatranja.

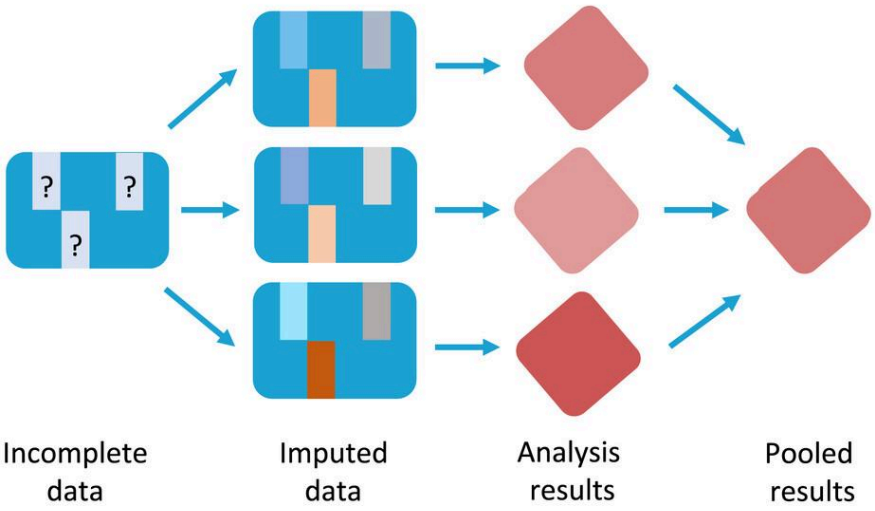


Slika 6. Primer LOCF imputacije

Važno je napomenuti da su ove metode pogodne kada instance ili posmatranja imaju vremenski ili sekvencijalni odnos. Međutim, ako su instance ili posmatranja nezavisna jedna od druge, kao što je slučaj kod specifičnih mehanizama nedostajanja podataka, ove metode možda nisu odgovarajuće.

3.2.2. Višestruka imputacija (Multiple Imputation)

Višestruka imputacija podrazumeva zamenu nedostajuće vrednosti sa više mogućih vrednosti, obično korišćenjem statističkih modela i algoritama. Višestruka imputacija generiše više kompletiranih datasetova, pri čemu svaki dataset sadrži imputirane vrednosti za nedostajuće podatke. Ovi datasetovi se zatim analiziraju korišćenjem standardnih statističkih metoda, a rezultati se kombinuju kako bi se dobili validni zaključci i procene.



Slika 7. Višestruka imputacija

Ove metode imaju nekoliko prednosti u odnosu na metode jednostruke imputacije. Generisanjem više imputiranih datasetova, uzima u obzir neizvesnost povezanu sa imputacijom nedostajućih vrednosti. Takođe, obezbeđuje varijabilnost između imputacija, što rezultira tačnijim procenama. Postoji veliki broj metoda i algoritama za implementaciju ovog tipa imputacije. Neki od poznatih pristupa uključuju metode *Maksimalne Verovatnoće* (*Maximum Likelihood*) i *Bajesovske metode* (*Bayesian methods*). Ove metode se razlikuju po svojim pretpostavkama, tehnikama modeliranja i pristupu specifičnim tipovima podataka. Izbor metode zavisi od karakteristika dataset-a, tipa nedostajućih podataka i ciljeva istraživanja.

Maximum Likelihood Imputation (MLI)

Maksimalna verovatnoća imputacije (Maximum Likelihood Imputation) je pristup zasnovan na modelu koji procenjuje nedostajuće vrednosti maksimizacijom funkcije verovatnoće u okviru specifičnog probabilističkog modela. Ova metoda tretira nedostajuće vrednosti kao latentne varijable i pronalazi vrednosti parametara koje čine da posmatrani podaci budu najverovatniji, uzimajući u obzir pretpostavljeni model. To podrazumeva formulisanje funkcije verovatnoće kao proizvoda uslovnih verovatnoća, pri čemu se uslovna verovatnoća svake nedostajuće vrednosti maksimizuje u odnosu na nepoznate parametre.

Metoda zasnovana na MLI obezbeđuje konzistentne imputirane vrednosti koje su u skladu sa pretpostavljenim statističkim modelom i posmatranim podacima. Međutim, oslanja se na specifične pretpostavke o raspodeli podataka, pa ako model nije dobro prilagođen podacima, imputirane vrednosti mogu biti pristrasne. Pored toga, MLI može biti osetljiv na izuzetke (outliers) i možda neće optimalno funkcionisati sa složenim obrascima nedostajanja podataka ili podacima sa velikim brojem dimenzija.

- **Expectation Maximization (EM)**

Jedna od tehnika unutar metode MLI je EM (Expectation Maximization) algoritam. EM algoritam je iterativni algoritam koji ima za cilj da pronađe procene maksimalne verovatnoće i prilagodi modele problemima sa nedostajućim podacima. Umesto direktne imputacije nedostajućih vrednosti, EM algoritam procenjuje podatke kroz dva koraka: korak očekivanja (E-Step) i korak maksimizacije (M-Step). Ovaj proces se ponavlja sve dok se ne dobiju procene maksimalne verovatnoće imputacije. U koraku očekivanja, procenjuju se nedostajuće vrednosti na osnovu trenutnih procena parametara modela, dok se u koraku maksimizacije parametri modela ažuriraju kako bi maksimizovali verovatnoću datih podataka. Proces se ponavlja dok procene ne konvergiraju, čime se postiže konačna imputacija podataka.

EM algoritam može se koristiti za procenu srednjih vrednosti, standardne devijacije i korelacija od interesa. Međutim, EM ima određena ograničenja. Zahteva veliki uzorak i pretpostavlja da je tip nedostajućih podataka MAR. Iako je konvergencija zagarantovana pod MAR pretpostavkom, brzina konvergencije zavisi od procenta nedostajućih podataka. Nizak procenat nedostajanja vodi do brze konvergencije, dok visok procenat usporava konvergenciju. Pored toga, EM je složena metoda, a njegova konvergencija može biti spora i može dovesti do suboptimalnih rezultata. Stoga je EM dobar algoritam za rešavanje problema sa MAR podacima, ali nije efikasan za MNAR podatke.

Bayesian metode

U Bajesovskim modelima imputacije nedostajućih podataka, nedostajuće vrednosti se tretiraju kao nepoznati parametri nasumično izvučeni iz odgovarajuće raspodele verovatnoće. Bajesovska paradigma pruža prirodan i fleksibilan okvir za modelovanje neizvesnosti u vezi sa nedostajućim vrednostima, procenjujući njihovu posterior raspodelu na osnovu posmatranih podataka i bilo koje dostupne prethodne informacije. Proces Bajesovske imputacije uključuje specifikaciju probabilističkog modela koji opisuje odnos između posmatranih podataka i nedostajućih vrednosti.

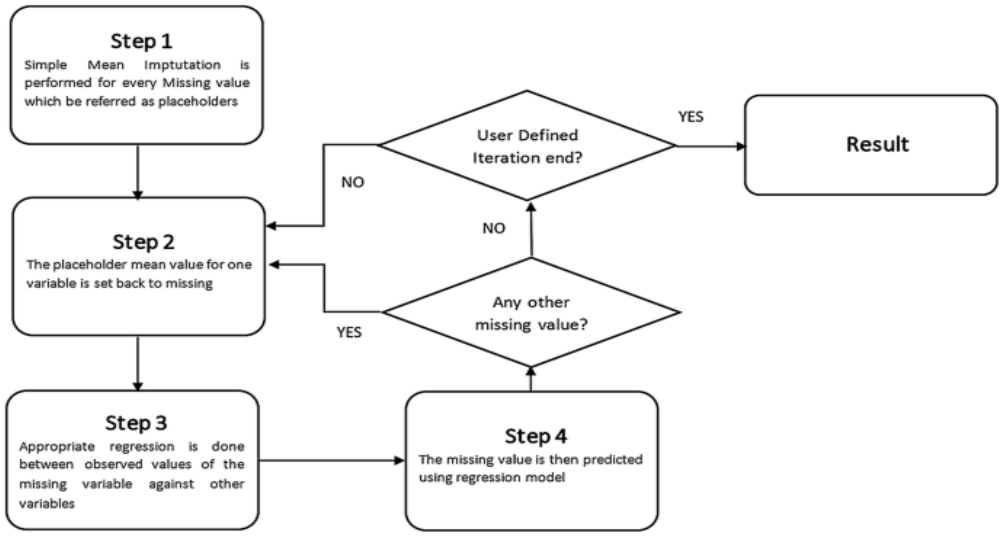
Ovaj model može da sakupi različite izvore informacija, poput prethodnih verovanja o nedostajućim vrednostima ili osnovnog procesa generisanja podataka. Kombinovanjem posmatranih podataka sa probabilističkim modelom, Bajesovske metode inferuju verovatne vrednosti nedostajućih podataka kroz procenu posterior raspodele.

- **Multiple Imputation by Chained Equation (MICE)**

Multiple Imputation by Chained Equations (MICE) je fleksibilna i moćna tehnika za imputaciju nedostajućih podataka koja je postala široko prihvaćena u različitim domenima analize podataka. MICE algoritam se koristi za rešavanje problema kada nedostajuće vrednosti nisu jednostavno slučajne, a posebno je efikasan u slučajevima kada podaci prate mehanizam koji je zavistan od drugih promenljivih u datasetu, odnosno MAR. Ovaj algoritam radi kroz iterativni proces gde se za svaku promenljivu sa nedostajućim vrednostima kreira model na osnovu drugih dostupnih varijabli.

Proces lančane jednačine može se razložiti na nekoliko opštih koraka:

1. Za svaki podatak koji nedostaje u skupu podataka, vrši se jednostavna imputacija, poput imputacije sa srednjom vrednošću. Ove srednje vrednosti mogu se smatrati *placeholder*-ima.
2. *Placeholder* imputacija jedne varijable se opet postavlja na nedostajuću..
3. Uočene vrednosti varijable u koraku 2 se regresiraju na druge varijable u populacionom modelu, koje mogu, ali ne moraju uključivati sve atribute u skupu podataka. Drugim rečima, nedostajuća vrednost je zavisna varijabla u regresionom modelu, dok su sve ostale varijable nezavisne. Ovi regresioni modeli funkcionišu pod istim pretpostavkama kao kada koristite linearne ili logističke regresione modele van konteksta popunjavanja podataka koji nedostaju.
4. Nedostajuće vrednosti se zatim zamenjuju predikcijama (vrednostima) iz regresionog modela. Kada se ta varijabla kasnije koristi kao nezavisna promenljiva u regresionim modelima za druge varijable, koristiće se i posmatrane vrednosti i ove popunjene vrednosti.
5. Koraci 2-4 se ponavljaju za svaku varijablu kojoj nedostaju podaci. Prolazak kroz svaku varijablu čini jednu iteraciju ili "ciklus". Na kraju jednog ciklusa, sve nedostajuće vrednosti se zamenjuju predikcijama iz regresija koje reflektuju odnose zabeležene u podacima.
6. Koraci 2-4 se ponavljaju kroz nekoliko ciklusa, a popunjeni podaci se ažuriraju nakon svakog ciklusa. Možemo odrediti broj ciklusa koje treba izvesti. Nakon ovih ciklusa, konačne popunjene vrednosti se zadržavaju, čime se dobija jedan kompletiran skup podataka. Obično se sprovodi 10 ciklusa, ali je potrebno istraživanje kako bi se odredio optimalan broj ciklusa za popunjavanje podataka u različitim uslovima. Ideja je da na kraju ciklusa distribucija parametara koji regulišu popunjavanje (npr. koeficijenti u regresionim modelima) konvergira ka stabilnosti. Ovo će, na primer, izbeći zavisnost od redosleda varijabli za popunjavanje.



Slika 8. MICE algoritam

- **Markov Chain Monte Carlo (MCMC)**

Pored MICE, tu je i Markov Chain Monte Carlo (MCMC). MCMC je moćan algoritam koji se koristi za imputaciju nedostajućih podataka, posebno u situacijama kada je problem nedostajućih podataka složen i multivarijantan. MCMC aproksimira zajedničku posterior raspodelu kada je evaluacija te prave raspodele analitički teška ili nemoguća zbog arbitrarnih obrazaca nedostajućih vrednosti i različitih tipova podataka (kontinuirani, nominalni, binarni, ordinalni).

MCMC algoritam funkcioniše kroz iterativne I-korake i P-korake. Na svakoj iteraciji, I-korak vuče imputacije iz prediktivne raspodele tekuće iteracije, uzimajući u obzir obrazac nedostajućih promenljivih za svaki slučaj. P-korak ažurira vrednosti parametara za prediktivnu raspodelu uzimajući uzorke iz posteriorne raspodele kompletiranih podataka. MCMC algoritam konvergira nakon dovoljno iteracija, pružajući imputacije za nedostajuće vrednosti koje simuliraju uzorke iz prave zajedničke posterior raspodele.

MCMC imputacija je zasnovana na Bajesovskim računarskim algoritmima, omogućavajući procenu srednje vrednosti, varijanse, kovarijacione matrice i drugih esencijalnih elemenata potrebnih za imputaciju. Ova metoda uzima u obzir varijabilnost podataka, koristi sve dostupne podatke i pruža imputirane vrednosti kao početne tačke za dopunjavanje nedostajućih vrednosti. Međutim, nedostaci uključuju pretpostavku multivarijantne normalne raspodele, visoke računarske zahteve i potrebu za velikim brojem iteracija.

3.2.3. Metode zasnovane na mašinskom učenju (Machine learning based methods)

Metode imputacije zasnovane na mašinskom učenju koriste metode nadgledanog ili nenadgledanog učenja kako bi procenile nedostajuće vrednosti u datasetovima, oslanjajući se na dostupne informacije iz podataka koji nisu nedostajući za preciznije predikcije. Prednost mašinskog učenja leži u njegovoj prediktivnoj sposobnosti da uhvati složene odnose i obrasce unutar podataka. Ove metode su fleksibilne, otporne na šum i izuzetke, i mogu da obrade različite tipove podataka, prilagođavajući se različitim obrascima nedostajućih podataka.

Njihova efikasnost u smanjenju pristrasnosti i obradi velikih datasetova čini ih moćnim i svestranim rešenjem za povećanje tačnosti i pouzdanosti analiza koje uključuju nepotpune podatke. Neki od uobičajenih pristupa su regresija, klasifikacija i klasterizacija.

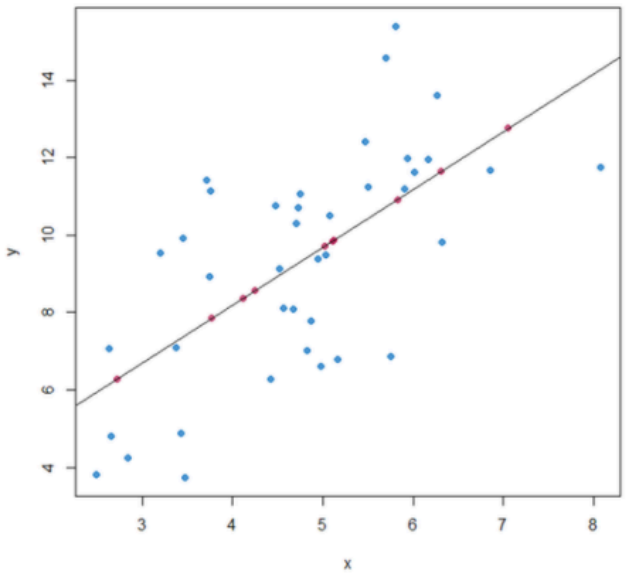
- **Regresiona imputacija (Regression Imputation)**

Regresija je metoda nadgledanog učenja. Metode imputacije nedostajućih podataka zasnovane na regresiji koriste regresione modele za procenu nedostajućih vrednosti u tabelarnim datasetovima. Ove metode prave model koristeći posmatrane podatke, tretirajući promenljivu sa nedostajućim vrednostima kao zavisnu promenljivu, dok druge kompletne promenljive koriste kao prediktore. Regresioni model se zatim koristi za predikciju nedostajućih vrednosti na osnovu vrednosti prediktorskih promenljivih.

Pretpostavimo da je X_{ik} nedostajuća vrednost u *k*-toj koloni i-tog primera. Linearni regresioni model za imputaciju će napraviti model prikazan ispod:

$$X_{ik} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik}$$

Standardne tehnike imputacije zasnovane na regresiji uključuju imputaciju prosekom/modom, jednostavnu linearnu regresiju, višestruku imputaciju i nelinearnu regresiju. Dok su linearna i logistička regresija pogodne za datasetove sa linearnim odnosima između promenljivih i mogu da obrade kontinuirane i kategorijske podatke, one možda nisu idealne za datasetove sa složenim ili nelinearnim obrascima. Tačnost imputiranih vrednosti zavisi od izbora prediktorskih promenljivih i performansi regresionog modela, zbog čega je pažljiv izbor od suštinskog značaja za pouzdane imputacije.



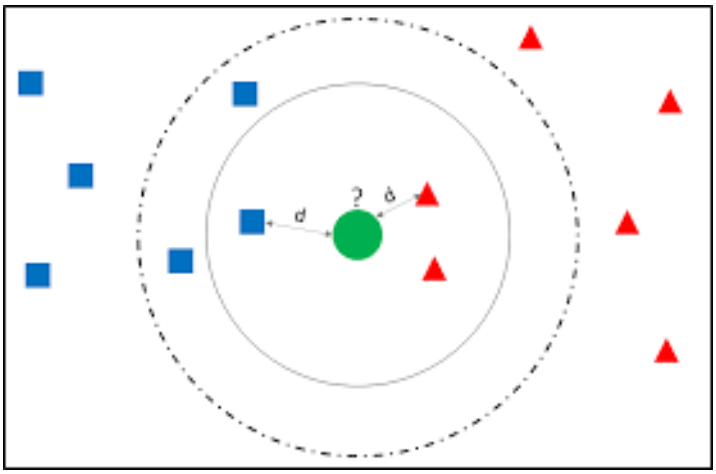
Slika 9. Primer regresione imputacije

Međutim, u nekim slučajevima, regresioni model nije efikasan, jer je potrebno ponovo prilagoditi model svaki put kada se promene delovi podataka sa nedostajućim vrednostima i posmatranim podacima.

- **KNN imputacija (Imputacija pomoću k najbližih suseda)**

K-nearest neighbour (K-NN) imputacija, popularna metoda nadgledanog učenja, može se koristiti i za imputaciju nedostajućih vrednosti u datasetovima. U ovom slučaju, vrednosti će biti popunjene srednjom vrednošću od k najbližih sličnih kompletnih opservacija. Sličnost između dve opservacije se određuje, nakon normalizacije skupa podataka, korišćenjem funkcije rastojanja koja može biti Euklidska, Manhetnova, Mahalanobisova, Pirsonova itd. Glavna prednost kNN algoritma je što, uz dovoljno podataka, može sa dobrom tačnošću predvideti uslovnu verovatnoću oko određene tačke i na taj način doneti dobro informisane procene. Može

predvideti diskretne i kontinuirane atribute. Još jedna prednost ove metode je što se uzima u obzir korelaciona struktura podataka. Izbor k vrednosti je veoma kritičan. Veća vrednost k bi uključila atribute koji se značajno razlikuju od ciljnog posmatranja, dok bi manja vrednost k mogla da izostavi značajne atribute.



Slika 10. Primer KNN

U slučajevima kada mehanizam nedostajućih podataka sledi MAR princip i ne postoji prethodno znanje o njihovoj raspodeli, ali se informacije mogu prikupiti iz posmatranih podataka, K-NN imputacija postaje pogodan izbor. Međutim, K-NN možda nije najprikladnija metoda za MNAR slučajeve, gde nedostajanje podataka nije povezano sa posmatranim podacima. Takođe, može biti dosta računarski zahtevna kod velikih skupova podataka.

- **Support Vector Machine (SVM)**

SVM je često korišćen algoritam mašinskog učenja za imputaciju nedostajućih podataka. Cilj SVM-a je da pronađe optimalnu razdvajajuću hiperravan u označenom skupu podataka, maksimizujući razdaljinu između hiperravni i najbližih tačaka podataka.

U kontekstu imputacije nedostajućih podataka, SVM može biti korišćen kao regresor ili klasifikator kako bi se predvidele vrednosti koje nedostaju. Na primer, regresioni SVM se koristi za imputaciju numeričkih podataka, dok se SVM klasifikator koristi za kategorijske podatke.

Jedna od prednosti SVM-a je sposobnost da identifikuje složene granice između klasa ili vrednosti u podacima, što mu omogućava da precizno proceni nedostajuće vrednosti u složenim skupovima podataka. Međutim, SVM za imputaciju nedostajućih podataka može biti osetljiv na distribuciju podataka i zahteva pažljivo podešavanje hiperparametara, kao što su izbor kernela i parametar regularizacije.

SVM je efikasan za imputaciju podataka koji prate MAR mehanizam, ali može imati ograničenja kada je u pitanju MNAR mehanizam.

- **Klasterovanje**

Klasterovanje, tehnika nenadgledanog učenja, grupiše slične elemente zajedno na osnovu funkcija sličnosti ili distance. Uobičajene metode klasterovanja, kao što je k-means klasterovanje, istražene su u brojnim studijama za obradu nedostajućih podataka. K-means metoda uključuje nasumično dodeljivanje centroida i iterativno preraspodeljivanje tačaka podataka ka najbližim centroidima kako bi se formirali klasteri. Ovaj proces se nastavlja sve dok dodeljivanje ne postane stabilno, a zatim se informacije o klasterima koriste za obradu nedostajućih vrednosti.

Važno je pažljivo razmotriti izbor metoda klasterovanja i njihovih parametara kada se klasterovanje primenjuje za imputaciju, jer rezultati u velikoj meri zavise od ovih izbora. Kako se veličina i dimenzionalnost dataset-ova rapidno povećavaju, tradicionalne metode imputacije zasnovane na statistici i mašinskom učenju mogu se suočiti sa izazovima u obradi podataka velikih razmera i visoke dimenzionalnosti. Pored toga, rastuća složenost i raznovrsnost formata i tipova podataka mogu dodatno ograničiti primenjivost ovih metoda.

- **Deep Learning metode**

Metode dubokog učenja za imputaciju nedostajućih podataka predstavljaju napredne pristupe koji koriste duboke neuronske mreže kako bi predvidele vrednosti koje nedostaju u velikim i kompleksnim skupovima podataka. Ovi modeli imaju sposobnost da automatski uče složene nelinearne odnose između varijabli, čineći ih posebno pogodnim za imputaciju u situacijama gde tradicionalne statističke ili mašinsko-učeće metode ne uspevaju da uhvate dublje obrasce u podacima. Neke od ključnih metoda dubokog učenja za imputaciju nedostajućih podataka su:

1. **Autoenkoderi**

Autoenkoderi su vrsta neuronske mreže koja uči da rekonstruiše svoje ulaze. U kontekstu imputacije, mreža se obučava da rekonstruiše kompletne podatke iz nepotpunih ulaza, koristeći kodirani (latentni) prostor kako bi „naučila“ relevantne obrasce u podacima. Ovaj latentni prostor pomaže u predikciji vrednosti koje nedostaju. Autoenkoderi su korisni za velike, visoko-dimenzionalne podatke i mogu da uhvate nelinearne odnose između varijabli.

2. **Generative Adversarial Networks (GANs)**

Generativne Adversarijalne mreže se sastoje od dve mreže: generatora i diskriminatora, koje se međusobno takmiče. U imputaciji, generator predlaže vrednosti koje nedostaju, dok diskriminator ocenjuje koliko su te vrednosti realistične. GAN-ovi se koriste za generisanje verodostojnih vrednosti koje nedostaju u podacima, posebno u slučajevima sa složenim distribucijama podataka. Ova metoda može kreirati imputacije koje izgledaju slično originalnim podacima.

3. **Recurrent Neural Networks (RNNs) i Long Short-Term Memory (LSTM)**

RNN-ovi, posebno LSTM varijante, koriste se za imputaciju u vremenskim serijama. Oni imaju sposobnost da se nose sa sekvencijalnim podacima i mogu predviđati vrednosti koje nedostaju na osnovu prethodnih i budućih vrednosti u nizu. Zbog svoje memorijske arhitekture, LSTM-ovi su veoma efikasni u hvatanju dugoročnih zavisnosti u vremenskim serijama, omogućavajući precizniju imputaciju.

Deep learning metode mogu da prepoznaju složene nelinearne odnose u podacima i rade sa velikim i kompleksnim datasetovima. Mogu da se prilagode različitim tipovima podataka, od numeričkih i kategorijalnih do vremenskih serija i slika. Međutim, ove metode zahtevaju velike količine podataka i resursa za treniranje. Često su potrebni napredni pristupi za podešavanje hiperparametara, što može biti izazovno.

3.3. Modeli mašinskog učenja otporni na nedostajuće podatke

Modeli mašinskog učenja koji su otporni na nedostajuće podatke mogu efikasno da se nose sa nepravilnostima i nedostacima u podacima, minimizujući uticaj nedostajućih vrednosti na tačnost i pouzdanost modela.

3.3.1. Decision Trees (Stabla odluke)

Decision Trees su fleksibilni modeli mašinskog učenja koji mogu prirodno da se nose sa nedostajućim podacima. Pri konstrukciji stabla, algoritam koristi karakteristike podataka kako bi kreirao čvorove koji dele skup podataka na podskupove, a pravila podele su zasnovana na merama poput entropije, Gini indeksa ili informacionog dobitka.

Kada je reč o nedostajućim podacima, Decision Trees omogućavaju da se ti podaci uključe u proces podele bez potrebe za prethodnom imputacijom. Algoritam može da napravi posebne grane za podatke koji nedostaju, tretirajući ih kao poseban uslov u procesu podele. Na primer, ako nedostaje vrednost za određenu karakteristiku, algoritam može da odluči kako će podaci biti podeljeni na osnovu drugih dostupnih karakteristika. Ova fleksibilnost omogućava modelima da budu otporni na nedostajuće podatke i da koriste sve dostupne informacije u izgradnji stabla. Pored toga, tehnike poput Random Forests, koje kombinuju više stabala odlučivanja, mogu dodatno smanjiti uticaj nedostajućih podataka, jer se svako stablo može nositi sa nedostajućim vrednostima na svoj način.

Međutim, u slučaju visokog stepena nedostajućih podataka ili složenih obrazaca nedostajanja, performanse stabla mogu biti smanjene, pa se u nekim slučajevima može primeniti imputacija nedostajućih vrednosti pre primene stabla odlučivanja.

Metode zasnovane na stablima pokazuju dobru performansu u obradi MCAR i MAR vrednosti i mogu se u određenoj meri nositi i sa MNAR vrednostima. Pored toga, nedostajući podaci i izuzeci (outliers) imaju minimalan uticaj na algoritme stabla odluke.

Imputation Method	Numerical	Categorical	MCAR	MAR	MNAR
Listwise Deletion	✓	✓	✓	×	×
Pairwise Deletion	✓	✓	✓	×	×
Mean/Median	✓	×	✓	×	×
Mode	×	✓	✓	×	×
LOCF & NOCB	✓	✓	✓	×	×
Maximum Likelihood	✓	×	✓	✓	×
Matrix Completion	✓	✓	✓	✓	✓
Bayesian Approach	✓	×	✓	✓	✓
Regression	✓	×	✓	✓	×
K-Nearest Neighbour	✓	✓	✓	✓	×
Tree Based	✓	✓	✓	✓	✓
SVM Based	✓	×	✓	✓	✓
Clustering Based	✓	✓	✓	×	×

Slika 11. Pregled metoda za imputaciju i njihovih osobina

Slika 11 pruža pregled različitih metoda imputacije zasnovanih na statistici i mašinskom učenju, zajedno sa odgovarajućim tipovima podataka i mehanizmima nedostajućih podataka za svaku od metoda.

4. Zaključak

U seminarskom radu obrađen je problem nedostajućih podataka, kao i pregled osnovnih tipova i mehanizama nedostajućih podataka. Razmotrene su različite metode za rešavanje ovih problema, kao što su brisanje nedostajućih podataka, jednostruke i višestruke imputacione metode i naprednije metode dubokog učenja.

Nedostajući podaci su problem u svim oblastima istraživanja, kako zbog sve veće količine podataka, transfera i konverzije podataka, tako i zbog ljudskih grešaka i propusta. Kada se rešava problem nedostajućih podataka, ne postoji jedan metod koji bi bio pogodan za sve tipove podataka. Izbor modela zavisi od prirode podataka, tipa nedostajućih vrednosti, i specifičnih zahteva analize. Dok neki modeli mogu automatski da se nose sa nedostajućim vrednostima, drugi možda zahtevaju predhodnu imputaciju ili obrada podataka. Jednostavne metode mogu biti dovoljno dobre za MCAR podatke, dok MAR i MNAR zahtevaju naprednije pristupe.

Ono što se može izvući kao zaključak je da izbor prave metode za rešavanje nedostajućih podataka zavisi od konteksta, vrste nedostajućih podataka i ciljeva analize.

5. Literatura

1. Youran Zhou, Sunil Aryal. *A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms*
2. Garret M. Fitzmaurice, Michael G. Kenward, Geert Molenberghs, Anastasios A. Tsiatis, Geert Verbeke. *Handbook of Missing Data*
3. Salvador García, Julián Luengo, Francisco Herrera. *Data Preprocessing in Data Mining*
4. https://www.researchgate.net/publication/308007055_Missing_Data