Univerzitet u Nišu

Elektronski fakultet

Predmet: Prikupljanje i predobrada podataka za mašinsko učenje

Nedostajući podaci

Seminarski rad

Smer: Veštačka inteligencija i mašinsko učenje

Student:

Milica Spasić 1588

Profesor:

Doc. dr Aleksandar Stanimirović

Niš, semptembar 2024. godine

Sadržaj

1. Uvod

Danas podaci predstavljaju temelj za donošenje ključnih odluka i razvoj različitih aplikacija, posebno u oblastima poput statistike, mašinskog i dubokog učenja i analize podataka. U realnosti podaci nisu uvek potpuni. Nedostajući podaci mogu nastati zbog različitih razloga – od tehničkih problema prilikom prikupljanja, preko ljudskih grešaka, pa sve do namernih propusta ili usled nedostatka odgovora ispitanika u istraživačkim studijama. Ovakvi propusti u podacima mogu značajno otežati analizu, narušiti tačnost modela i dovesti do pogrešnih zaključaka.

Nedostajući podaci su čest izazov koji zahteva pažljivo razmatranje prilikom analize. Oni mogu negativno uticati na statističku analizu, dovesti do smanjenja pouzdanosti rezultata i uvesti nepostojeću pristrasnost. Većina tradicionalnih i savremenih statističkih i mašinskih metoda podrazumeva da su podaci kompletni, što dovodi do problema kada se suočimo sa nepotpunim opservacijama. U takvim situacijama, nepravilno rukovanje nedostajućim podacima može rezultirati gubitkom važnih informacija i smanjenjem statističke snage. Ono što je neophodno razumeti i primeniti jeste da su razumevanje prirode nedostajućih podataka i primena adekvatnih tehnika za njihovo rešavanje ključni koraci u bilo kakvoj analizi nad njima.

U ovom radu se pruža sveobuhvatan pregled problema nedostajućih podataka, analiza faktora koji dovode do njihovog nastanka, kao i metoda koje se koriste za njihovo rešavanje, a razmatramo prednosti i mane svake metode kako bismo naglasili ključne faktore koje treba uzeti u obzir pri izboru metode za upravljanje nedostajućim podacima u konkretnom istraživanju.

2. Tipovi nedostajućih podataka

Kada radimo sa datasetovima sa nedostajućim podacima, potrebno je utvrditi mehanizam po kojem podaci nedostaju, u cilju precizne analize podataka.

Postoje tri osnovna mehanizma nedostajanja podataka koji se mogu opisati u zavisnosti od odnosa između dostupnih i nedostajućih podataka:

1. MCAR (Missing Completely At Random)
2. MAR (Missing At Random)
3. MNAR (Missing Not At Random)

2.1. MCAR

Ovaj mehanizam nedostajanja podataka podrazumeva da se podaci pojavljuju nasumično i to bez ikakve korelacije sa drugim atibutima ili vrednostima koje nedostaju. Podatak je nedostajući ako verovatnoća da on bude nedostajući zavisi samo od samog podatka, a ne zavisi nikako od vrednosti drugih atributa ili nedostajućih vrednosti.

Primer iz svakodnevnog života za ovaj mehanizam može biti situacija u kojoj se prikupljaju odgovori putem online ankete, ali usled tehničkih problema, poput privremenog pada sistema, neki odgovori nisu sačuvani. U ovom slučaju, verovatnoća da određeni odgovori nedostaju ne zavisi ni od samih odgovora ni od karakteristika učesnika. Drugim rečima, svi odgovori imaju jednaku šansu da budu izgubljeni zbog slučajne greške.

Takvi podaci se mogu jednostavno zanemariti bez većih posledica po analizu, jer njihov izostanak nije povezan ni sa jednim atributom u skupu podataka. Algoritmi mašinskog učenja koji pretpostavljaju da podaci nedostaju nasumično neće biti pristrasni u ovakvim situacijama. Ipak, postoji mogućnost da se u ovom procesu izgubi deo statističke snage, ukoliko je broj instanci (vrsti) sa nedostajućim podacima značajno veliki, što može uticati na preciznost modela.

Zbog ove nasumične prirode, MCAR podaci se mogu relativno lako tretirati bez uvođenja pristrasnosti u analizu brisanjem ili popunjavanjem jednostavnim metodama zamene procenjenom vrednošću na osnovu dostupnih informacija(imputacija), kao što su popunjavanje srednjom vrednošću, medianom, modalnom vrednošću. Međutim, iako MCAR podaci ne ugrožavaju tačnost rezultata, brisanje ili nepravilna obrada može smanjiti statističku snagu modela.

2.2. MAR

Ovaj mehanizam se odnosi na slučaj kada verovatnoća da neka vrednost nedostaje zavisi od drugih dostupnih podataka, a ne zavisi od od same vrednosti koja nedostaje. Dakle, postoji povezanost između nedostajućih podataka i dostupnih vrednosti i atributa, ali ne i sa samom vrednossću koja nedostaje. Ovo nam omogućava procenu nedostajuće vrednosti na osnovu postojećih podataka.

Primer iz svakodnevnog života za ovaj mehanizam može biti situacija u kojoj se prikupljaju podaci o zdravlju i težini ljudi, uključujući i njihov BMI (indeks telesne mase) i nivo fizičke aktivnosti. Pretpostavimo da ispitanici koji su manje fizički aktivni češće izostavljaju podatak o svojoj težini jer se možda osećaju nelagodno ili nesigurno. Međutim, činjenica da ne prijavljuju svoju težinu ne zavisi direktno od toga koliko zapravo teže, već od njihovog nivoa fizičke aktivnosti.

Dakle, u ovom slučaju, nedostajući podaci o težini su povezani sa posmatranom promenljivom – fizičkom aktivnošću, ali ne direktno sa vrednostima težine. To je primer MAR mehanizma, jer je moguće predvideti ili imputirati nedostajuće vrednosti težine na osnovu informacija o fizičkoj aktivnosti ispitanika, bez direktnog uvida u njihove stvarne težine. Mehanizam nedostajanja nije nasumičan, ali je dovoljno povezan sa drugim posmatranim podacima da može biti modelovan na osnovu njih.

Ovaj tip podataka omogućava primenu različitih tehnika imputacije, jer se nedostajuće vrednosti mogu predvideti na osnovu dostupnih informacija. Uprkos tome što MAR podaci nisu potpuno nasumični, njihovo pravilno tretiranje može značajno smanjiti pristrasnost i poboljšati tačnost analize. Zbog toga je važno pravilno prepoznati i koristiti odgovarajuće metode imputacije kako bi se nadomestili nedostajući podaci i osigurali pouzdani rezultati. Imputacija zasnovana na algoritmu kao što K najbližih suseda (k-NN) ili regresione imputacione metode mogu uspešno obraditi MAR podatke jer koriste imputacione tehnike koje uzimaju u obzir vrednosti ostalih varijabli.

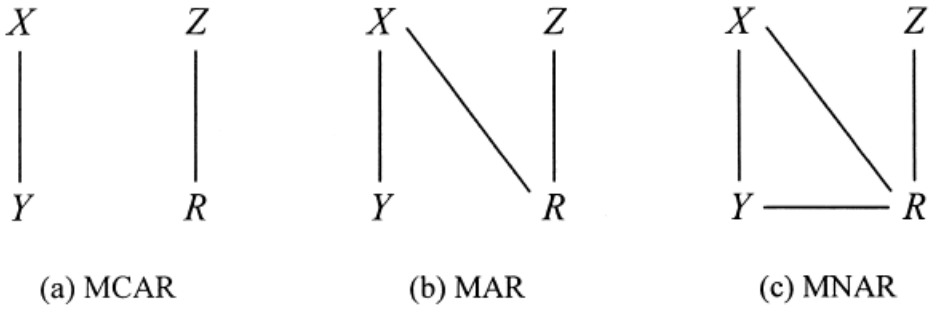
2.3. MNAR

Ovi podaci su tip nedostajućih podataka gde verovatnoća da neki podatak nedostaje zavisi od samih vrednosti koje nedostaju, odnosno podaci nisu nasumično odsutni. Drugim rečima, razlog za nedostajanje podataka leži u samim podacima, a ne u drugim posmatranim promenljivama.

Primer MNAR mehanizma može biti istraživanje o konzumiranju alkohola. Možemo da pretpostavimo da ljudi koji piju velike količine alkohola češće izbegavaju da odgovore na pitanja o tome zbog osećaja krivice ili straha od osude. U tom slučaju, podaci nedostaju upravo zato što te osobe piju više, a ne zbog nekih drugih faktora, poput godina ili pola.

Ovo je klasičan primer MNAR mehanizma, jer je nedostajanje podataka direktno povezano sa količinom alkohola koju osoba konzumira. Takvi podaci su teži za analizu jer nedostatak nije nasumičan, što otežava popunjavanje nedostajućih vrednosti, kao i preciznu analizu.

MNAR podaci predstavljaju jedan od najkompleksnijih izazova u analizi podataka jer verovatnoća da podaci nedostaju zavisi od samih vrednosti koje su odsutne. Upravo zbog te povezanosti, uobičajene metode imputacije često nisu pouzdane, jer ne mogu pravilno nadomestiti nedostajuće vrednosti bez uvođenja pristrasnosti. Da bi se pravilno analizirali MNAR podaci, potrebno je koristiti napredne tehnike ili modele koji uzimaju u obzir razloge zbog kojih podaci nedostaju, kako bi se smanjio uticaj pristrasnosti i obezbedila što tačnija analiza. Razumevanje prirode ovih podataka ključno je za uspešnu obradu i interpretaciju rezultata.



Slika 1: Tipovi nedostajućih podataka

Na Slici 1 prikazana su sva tri mehanizma nedostajućih podataka. Atribut X predstavlja posmatrane podatke, bez nedostajućih vrednosti, dok je atribut Y onaj kod kojeg podaci nedostaju. Z simbolizuje komponentu koja nije povezana ni sa jednim od ovih atributa, a R označava odsustvo podataka, odnosno mehanizam njihovog nedostajanja.

Svako istraživanje je specifično, pa je pre same analize ključno razumeti prirodu podataka i razmotriti sve faktore koji mogu uticati na njihovu pouzdanost i validnost. Bez obzira na te okolnosti, cilj istraživanja treba biti usmeren na primenu preporučenih pristupa u analizi i obradi nedostajućih podataka, kako bi se obezbedili pouzdani i validni rezultati.

3. Metode za obradu nedostajućih podataka

Postoji mnogo različitih pristupa za rukovanje nedostajućim podacima, koji se mogu podeliti na tradicionalne i savremene metode. Tradicionalni pristupi uključuju brisanje nedostajućih vrednosti, dok moderni pristupi koriste metode zasnovane na modelima, poput metode maksimalne verodostojnosti i višestruke imputacije. Ove tehnike mogu se primenjivati kako nad numeričkim, tako i nad kategoričkim atributima. Ipak, neke metode nisu najbolje prilagođene za kategoričke atribute, dok su druge posebno preporučene za takve slučajeve.

Metode za obradu nedostajućih podataka treba prilagoditi specifičnostima datog seta podataka, uzrocima nedostatka i procentu nedostajućih vrednosti. Uglavnom se biraju metode koje su jednostavne za primenu i koje minimalno uvode pristrasnost u podatke.

Kada su podaci MCAR ili MAR, razlozi zbog kojih podaci nedostaju mogu se zanemariti, što olakšava odabir metode, jer se bilo koja može primeniti. Međutim, često je teško sa sigurnošću utvrditi da li podaci spadaju u MCAR ili MAR kategoriju. Dobra strategija je ispitivanje osetljivosti rezultata kroz primenu različitih metoda i poređenje rezultata, kako bi se procenilo koje pretpostavke najbolje odgovaraju datim podacima.

Najčešće korišćene metode za obradu podataka sa nedostajućim vrednostima se mogu podeliti u tri glavne kategorije: i**s**ključivanje instanci sa nedostajućim vrednostima, metoda maksimalne verodostojnosti i imputacija nedostajućih vrednosti.

Isključivanje ili brisanje (deletion) podataka koji imaju nedostajuće vrednosti uključuje brisanje atributa koji imaju visoke nivoe praznina, odnosno koji imaju veliki procenat nedostajućih podataka u sebi. Ovaj pristup je jednostavan i efikasan kada je procenat nedostajućih podataka relativno nizak. Međutim, može dovesti do značajnog gubitka informacija, što može negativno uticati na kvalitet analize.

Druga metoda je metoda maksimalne verovatnoće, koja prvo procenjuje parametre modela koristeći potpune podatke, a zatim ih koristi za imputaciju nedostajućih vrednosti putem uzorkovanja. Ovaj pristup može pružiti preciznije imputacije, posebno kod velikih skupova podataka, i omogućava očuvanje informacija. Njegovo sprovođenje može biti složeno i zahtevati mnogo resursa.

Postoji i imputacija nedostajućih vrednosti, gde se one popunjavaju procenjenim vrednostima. U većini slučajeva, atributi nisu nezavisni, pa se istraživanjem odnosa među njima mogu odrediti nedostajuće vrednosti. Ovaj metod omogućava zadržavanje svih opservacija u skupu podataka, što može rezultirati kvalitetnim procenama ako su odnosi pravilno identifikovani. Međutim, ako imputacija nije dovoljno precizna, može dovesti do pristrasnosti i smanjenja varijabilnosti.

3.1. Isključivanje (brisanje) instanci sa nedostajućim vrednostima

Ova tehnika predstavlja najjednostavniji način za rešavanje problema nedostajućih podataka i podrazumeva uklanjanje celih redova ili atributa sa velikim brojem nedostajućih vrednosti. Iako je ova metoda jednostavna i efikasna, kada je veliki procenat nedostajućih podataka, može dovesti do značajnog gubitka informacija, što negativno utiče na kvalitet analize. Postoje isključivanje nedostajućih podataka u celini (Listwise deletion) i isključivanje nedostajućih podataka u parovima (Pairwise deletion). Ove metode brisanja daju validne rezultate samo u slučajevima MCAR podataka.

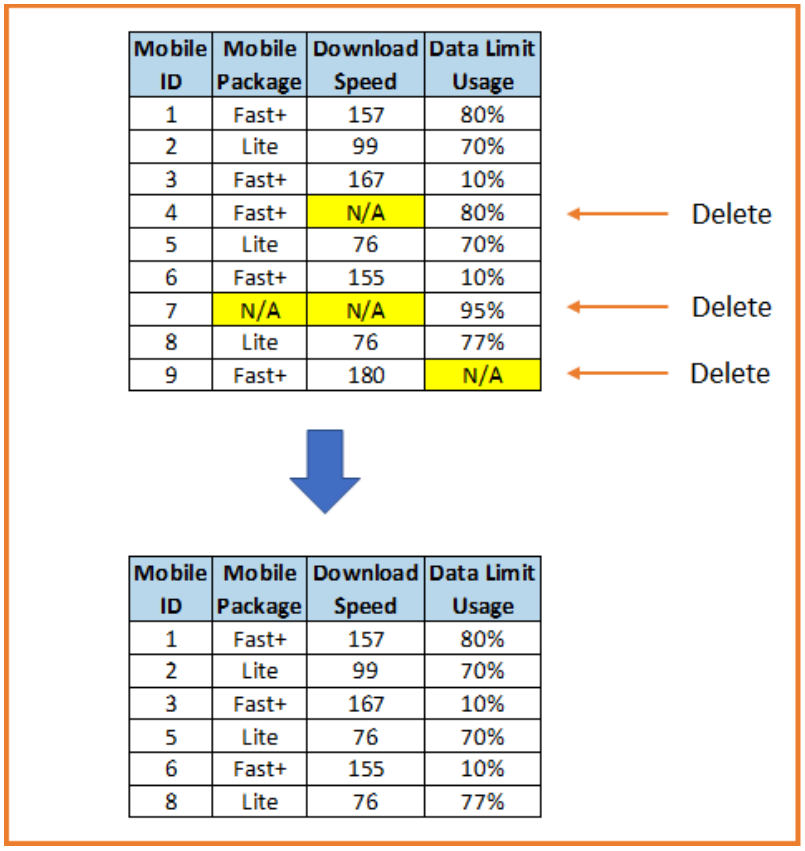
3.1.1. Listwise deletion

Listwise deletion (poznato i kao casewise deletion ili analiza potpunih slučajeva) je metoda za rukovanje nedostajućim podacima, u kojoj se iz dataset-a uklanjaju čitave instance (redovi) koji sadrže makar jednu nedostajuću vrednost. Dakle, ako bilo koji atribut u opservaciji ima nedostajuće podatke, ta opservacija se potpuno isključuje iz analize (Slika 2).

Ova metoda je vrlo jednostavna za implementaciju i ne zahteva dodatne korake, poput popunjavanja nedostajućih vrednosti ili razvijanja zamenskih rešenja. Najveća prednost listwise deletion metode leži u njenoj jednostavnosti, zbog čega se često koristi kada je procenat nedostajućih podataka mali.

Međutim, ukoliko procenat nedostajućih podataka postane visok, veliki broj opservacija može biti uklonjen, što može dovesti do smanjenja veličine uzorka i gubitka ključnih informacija. To dalje rezultira smanjenom statističkom snagom, jer manje opservacija znači veće standardne greške u procenama. Pored toga, ako podaci ne nedostaju potpuno nasumično (MCAR), listwise deletion može uvesti pristrasnost u analizu, jer preostali podaci možda neće biti verodostojna reprezentacija stvarne populacije.

Zbog ovih razloga, listwise deletion metoda je najbolje primenljiva kada su podaci MCAR i kada procenat nedostajućih vrednosti nije značajan, jer u suprotnom može ozbiljno narušiti kvalitet analize i pouzdanost rezultata.



Slika 2. Primer listwise brisanja

3.1.2. Pairwise deletion

Isključivanje nedostajućih podataka u parovima (eng. Pairwise deletion) je metoda za rukovanje nedostajućim podacima koja se razlikuje od listwise deletion po tome što ne uklanja cele redove podataka sa nedostajućim vrednostima. Umesto toga, za svaku pojedinačnu analizu koriste se samo opservacije koje imaju sve potrebne vrednosti za atribute uključene u tu specifičnu analizu. Na primer, ako neka opservacija ima nedostajuću vrednost u određenoj koloni, ona će biti isključena samo za analizu te kolone, ali će biti uključena u sve druge analize gde nedostajućih vrednosti nema (Slika 3).

Glavna prednost ove metode je što omogućava maksimalno korišćenje dostupnih podataka i smanjuje gubitak informacija, što je posebno korisno kada želite da zadržite što više podataka u analizi. Međutim, pairwise deletion može dovesti do nekonzistentnih rezultata jer različite analize koriste različite uzorke podataka. Ovo može stvoriti problem kod analize korelisanosti atributa, zato što bi se korelacije računale nad različitim skupovima podataka.

Jedan od glavnih nedostataka ove metode je to što, zbog različitih uzoraka podataka za različite analize, mogu biti generisane pristrasne procene parametara. Kako se ne koristi jedinstveni

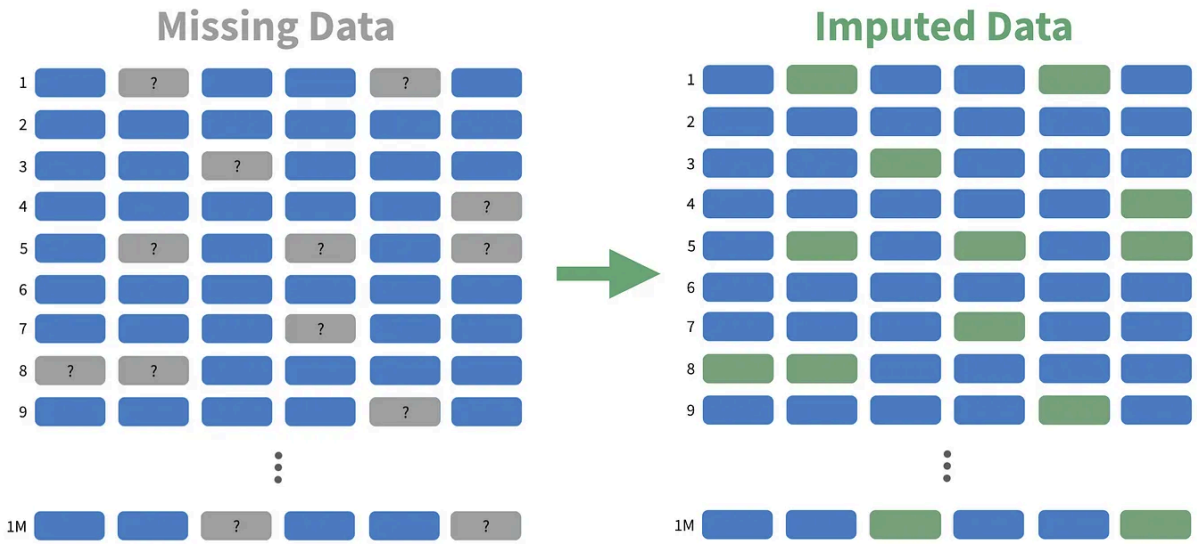
uzorak podataka za sve analize, ne postoji dosledna osnova za izračunavanje standardnih grešaka parametara, što može otežati interpretaciju rezultata i smanjiti preciznost analize.

Y	X_1	X_2	X_3
4	0.2	1.2	20
3	NA	1.2	21
2	0.3	1.1	16
2	0.4	1.1	17
1	0.5	2	18
2	0.4	2.1	18
NA	0.2	1.4	19
2	0.1	1.2	22
2	0.1	NA	NA

Slika 3. Primer pairwise brisanja

4. Imputacija podataka

Imputacija nedostajućih podataka je proces popunjavanja nedostajućih vrednosti u skupu podataka kako bi se osiguralo da su analize i modeli što potpuniji i tačniji (Slika 4). Ovo se može postići različitim metodama, kao što su zamena nedostajućih vrednosti srednjom, medijanom ili modom, korišćenje regresije za predviđanje nedostajućih vrednosti, ili upotreba složenijih tehnika kao što su K-najbližih suseda ili višestruka imputacija. Imputacija je važna jer nedostajući podaci mogu značajno uticati na rezultate analize, a pravilno popunjavanje tih praznina može poboljšati kvalitet i pouzdanost zaključaka koji se donose na osnovu skupa podataka.



Slika 4. Imputacija nedostajućih podataka

Metode imputacije se mogu podeliti u dve kategorije: jednostruku i višestruku imputaciju. Jednostruka imputacija se odnosi na proces popunjavanja nedostajućih vrednosti u skupu podataka koristeći samo jednu metodu ili procenu za svaku nedostajuću vrednost. Na primer, ako se koristi srednja vrednost, svaka nedostajuća vrednost se zamenjuje tom istom srednjom vrednošću. S druge strane, višestruka imputacija uključuje generisanje više različitih setova imputacija za svaku nedostajuću vrednost, obično koristeći različite metode ili uzorke, a zatim se rezultati kombinuju kako bi se dobila konačna procena. Ova metoda omogućava bolje procene neizvesnosti i smanjuje pristrasnost, jer uzima u obzir varijabilnost između različitih imputacija.

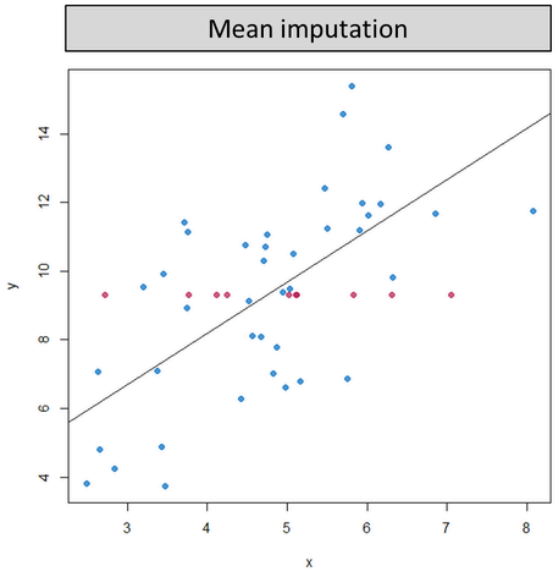
4.1 Jednostruka imputacija

Jednostruka imputacija predstavlja tehniku za popunjavanje nedostajućih vrednosti u skupu podataka, pri čemu se koristi jedna jedinstvena procena za svaku prazninu. Na primer, nedostajući podaci mogu biti zamenjeni srednjom vrednošću (Slika 5), medijanom ili onom vrednošću koja se najčešće pojavljuje u preostalim podacima. Još neke metode jednostruke imputacije su i LOCF(Last Observation Carried Forward), NOCB (Next Observation Carried Backward), imputacija korišćenjem regresije i slučajna imputacija. U nastavku ćemo ukratko obraditi svaku od njih.

4.1.1. Jednostruka imputacija korišćenjem srednje vrednosti, medijana, modusa

Ako imamo podatke o visinama, nedostajući unos može se zameniti prosečnom visinom (mean) preostalih osoba (Slika 5). Kada su podaci asimetrični ili imaju ekstremne vrednosti možemo izvršiti imputaciju medijanom, jer medijan nije pod uticajem tih ekstremnih podataka. Kada se radi o kategorički podacima, nedostajuće vrednosti mogu se popiniti modom, jer mode predstavlja najčešće pojavljivanje vrednosti u skupu podataka. Na primer, ako u anketi na neko pitanje najveći broj ljudi odgovori "da," ta vrednost se može koristiti za popunjavanje nedostajućih unosa.

Što se tiče analize numeričkih podataka, srednja vrednosti i medijan su prikladni, dok je mode efikasniji za kategoričke ili binarne podatke, koji ne poseduju srednju vrednost ili medijan. Treba imati na umu da, zbog svoje jednostavnosti, ove metode možda neće prikazati složenu strukturu raspodele podataka. Ove metode imputacije imaju određena ograničenja i efikasne su samo ako mehanizam nedostajanja podataka prati MCAR pretpostavku.

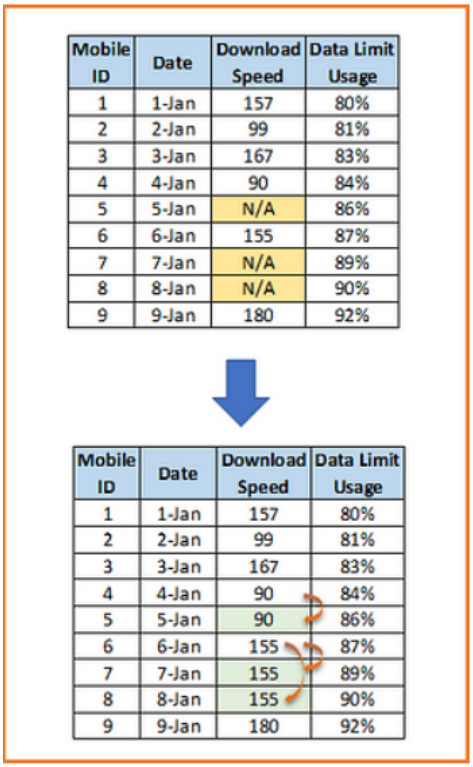


Slika 5. Ilustracija jednostruke imputacije korišćenjem srednje vrednosti

4.1.2. LOCF i NOCB

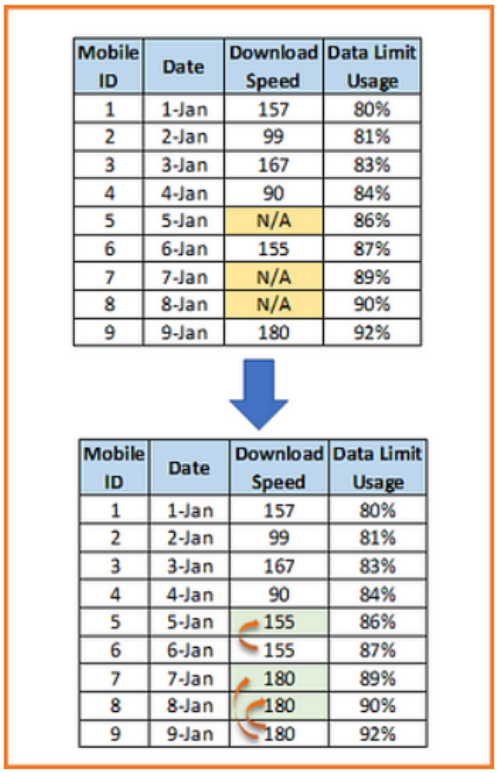
LOCF (Last Observation Carried Forward) i NOCB (Next Observation Carried Backward) su metode imputacije koje se koriste za popunjavanje nedostajućih vrednosti u vremenskim serijama ili sekvencijalnim podacima.

LOCF podrazumeva da se nedostajuća vrednost zameni poslednjom dostupnom vrednošću (Slika 6). Na primer, ako u seriji podataka nedostaje vrednost za određeni trenutak, koristi se vrednost iz prethodnog trenutka kako bi se popunila praznina. Ova metoda je jednostavna, ali može biti problematična jer može stvoriti pristrasnost ako se poslednje posmatrane vrednosti ne menjaju tokom vremena.



Slika 6. Primer LOCF imputacije

NOCB metoda funkcioniše na sličan način, ali umesto da koristi prethodnu vrednost, koristi se sledeća dostupna vrednost da bi se popunila nedostajuća pozicija. Dakle, ako nedostaje vrednost, zamenjuje se vrednošću koja dolazi nakon te praznine (Slika 7). Ova metoda takođe može dovesti do pristrasnosti ako se u podacima događaju značajne promene između posmatranja.



Slika 7. Primer NOCB imputacije

Obe metode su korisne u situacijama kada su podaci vremenski strukturirani, ali treba ih koristiti s oprezom zbog potencijalnog gubitka informacija o varijacijama u podacima.

4.1.3.Slučajna imputacija

Slučajna imputacija predstavlja metodu čija suština leži u zameni nedostajućih vrednosti slučajno odabranim vrednostima iz postojećih podataka. Ova tehnika je korisna u situacijama kada je važno zadržati prirodnu varijabilnost podataka, čime se omogućava da analize budu realističnije i bliže stvarnim okolnostima.

Proces slučajne imputacije počinje identifikacijom nedostajućih vrednosti u skupu podataka. Kada se utvrde mesta gde podaci nedostaju, pristupa se odabiru vrednosti za imputaciju. Umesto korišćenja neke centralne mere kao što su srednja vrednost ili medijan, nedostajuća vrednost se zamenjuje nasumično odabranom vrednošću iz iste varijable koja već postoji u skupu podataka. Na primer, ako u bazi podataka o visinama nedostaje određeni unos, može se odabrati slučajna visina iz postojećih podataka, čime se čuva varijabilnost.

Ukoliko se ne koristi pažljivo, može dovesti do pristrasnosti, slučajno favorizujući određene vrednosti koje možda ne odražavaju stvarnu raspodelu podataka. Takođe, ova metoda ne pruža

dodatne informacije o neizvesnosti u procenama nedostajućih vrednosti, što može smanjiti tačnost konačnih analiza.

Slučajna imputacija može biti korisna strategija u obradi nedostajućih podataka, ali je ključno razmotriti njene potencijalne posledice i koristiti je u kombinaciji s drugim metodama kako bi se obezbedila tačnost i validnost rezultata.

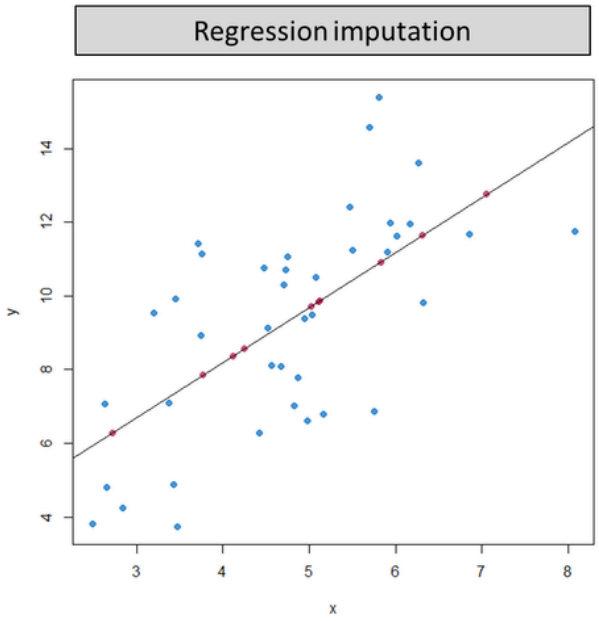
4.1.4. Imputacija korišćenjem regresije

Imputacija korišćenjem regresije najčešće podrazumeva linearnu regresiju, koja pomaže da se nedostajući podaci dobiju na osnovu drugih dostupnih informacija.

Prvo se identifikuju atributi koji nemaju podatke. Nakon toga, razvija se model koji koristi informacije iz drugih atributa kako bi predvideo nedostajuće vrednosti. Na primer, ako u podacima o potrošačima nedostaju informacije o prihodima, možemo iskoristiti atribute kao što su starost, obrazovanje ili mesto stanovanja kako bismo procenili te nedostajuće vrednosti.

Jedna od prednosti ove metode je što može da pokaže kako se varijable međusobno povezuju. Korišćenjem regresije, možemo bolje razumeti kako promene u jednoj varijabli utiču na druge. Ovo može dovesti do tačnijih procena nego što bi to bile jednostavne metode kao što su srednja vrednost. Kada se model obuči, koristi se za predikciju nedostajućih vrednosti koje se zatim unose u skup podataka (Slika 8).

Međutim, ako su pretpostavke o vezama između atributa netačne, to može dovesti do pogrešnih procena. Imputacija korišćenjem regresije može biti veoma korisna za popunjavanje nedostajućih podataka, posebno kada su dostupne dodatne informacije koje mogu pomoći u predikciji.



Slika 8. Ilustracija imputacije pomoću regresije

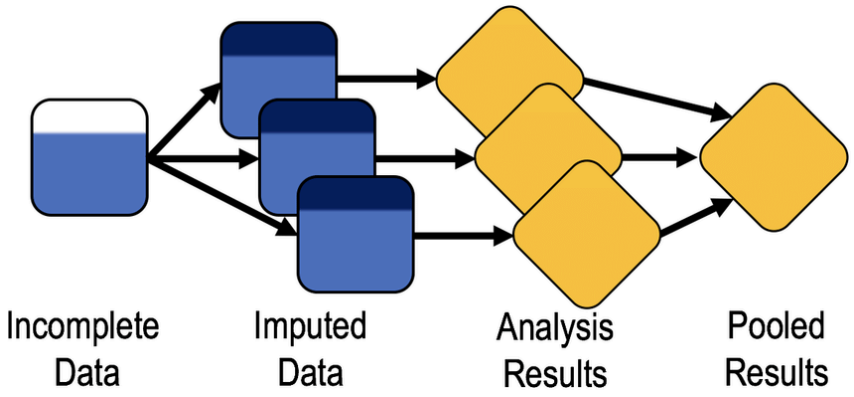
4.2 Višestruka imputacija

Višestruka imputacija (Multiple Imputation) je napredna statistička tehnika koja se fokusira na očuvanje informacija o varijabilnosti i neizvesnosti podataka, što je ključno za dobijanje tačnih i pouzdanih rezultata.

Kreiranje modela za imputaciju se vrši na osnovu dostupnih podataka, razvija se model koji opisuje vezu između različitih atributa. Generisanje više imputacija funkcioniše tako što višestruka imputacija generiše nekoliko (obično između 5 i 20) različitih setova nedostajućih vrednosti. Svaki set se stvara tako što se uzima u obzir varijabilnost u podacima. Na primer, ako su podaci o visini i težini, imputacija može koristiti raspodelu ovih varijabli da generiše različite, verovatne visine za nedostajuće unose.

Analiza svakog skupa funkcioniše tako da svaki od generisanih setova se analizira odvojeno koristeći iste statističke metode. To može uključivati regresijske analize, korelacije ili bilo koju drugu relevantnu analizu. Ove analize rezultiraju različitim procenama parametara, što odražava varijabilnost između setova.

Nakon što su analize izvršene, rezultati iz svih setova se kombinuju (Slika 9),



Slika 9. Vizualizacija procesa višestruke imputacije

Prednosti višestruke imputacije su to što umesto gubitka podataka ili jednostavnog popunjavanja vrednosti jednom procenom, višestruka imputacija čuva sve informacije i varijabilnost, kombinovanjem rezultata iz više setova, dobijaju se pouzdanije procene i validnija statistička procena. Ova tehnika može se koristiti u različitim kontekstima, uključujući vremenske serije, klinička ispitivanja i socijalne istraživačke podatke.

Implementacija višestruke imputacije može biti složena i zahteva više koraka u poređenju sa jednostrukim metodama imputacije, a tačnost rezultata zavisi od kvaliteta modela koji se koristi za imputaciju. Višestruka imputacija može zahtevati više računarskih resursa i vremena, posebno kada se radi sa velikim skupovima podataka.

4.2.1. Maksimalna verovatnoća imputacije (MLI)

Imputacija korišćenjem maksimalne verovatnoće (Maximum Likelihood Imputation) je jedna od naprednijih metoda koja se oslanja na statističke modele kako bi se na osnovu postojećih podataka procenile vrednosti koje nedostaju. Osnovna ideja je da se koristi raspodela podataka kako bi se maksimalizovala verovatnoća da su podaci koje imamo tačni, a nedostajuće vrednosti logične i konzistentne.

Proces počinje izgradnjom modela koji opisuje kako su različiti atributi u skupu podataka međusobno povezani. Na primer, ako radimo sa podacima o visini i težini, možemo pretpostaviti da visina i težina imaju određeni odnos. Statistički model se može prilagoditi tim podacima, a na osnovu tog modela izračunavamo parametre koji najbolje objašnjavaju raspodelu podataka.

Jednom kada se parametri modela odrede, koristi se stohastički pristup za predviđanje nedostajućih vrednosti. Ovo znači da se uzima u obzir varijabilnost podataka, pa se nedostajuće vrednosti ne predviđaju kao fiksne, već se generišu kao niz potencijalnih vrednosti koje su u skladu s modelom. Na primer, umesto da jednostavno zamenimo nedostajuću težinu prosečnom

težinom, možemo koristiti model da bismo stvorili niz vrednosti koje odražavaju verovatne težine na osnovu visine i drugih varijabli.

Jedna od glavnih prednosti imputacije korišćenjem maksimalne verovatnoće je tačnost procena koje se dobijaju, posebno u velikim i složenim skupovima podataka. Ova metoda može da uhvati složene odnose između atributa, čime se povećava verovatnoća da će dobijene imputacije biti realistične i korisne za analizu. Međutim, implementacija maksimalne verovatnoće može biti kompleksna i zahtevati naprednije statističke veštine.

4.2.2. Bayesian metoda

Bayesian metoda je pristup statistici koji se koristi za procenu verovatnoće nekih događaja ili parametara na osnovu prethodnog znanja, kao i novih informacija. Ova metoda se oslanja na Bayesovu teoremu, koja objašnjava kako ažurirati verovatnoće kada se dobiju novi podaci, odnosno nova saznanja.

Osnovni koncept Bayesove teoreme jeste uslovna verovatnoća. To znači da, kada imamo nove dokaze, možemo izračunati verovatnoću nekog događaja uzimajući u obzir te dokaze. Ova metoda omogućava istraživačima da kombinuju informacije iz različitih izvora i postepeno menjaju svoje verovatnoće, ili uverenja, kako dolaze novi podaci. Već poznate verovatnoće nazivaju se prior, a one koje dobijamo sa novim iskustvom, odnosno podacima, posterior.

4.2.3. Multiple Imputation by Chained Equations (MICE)

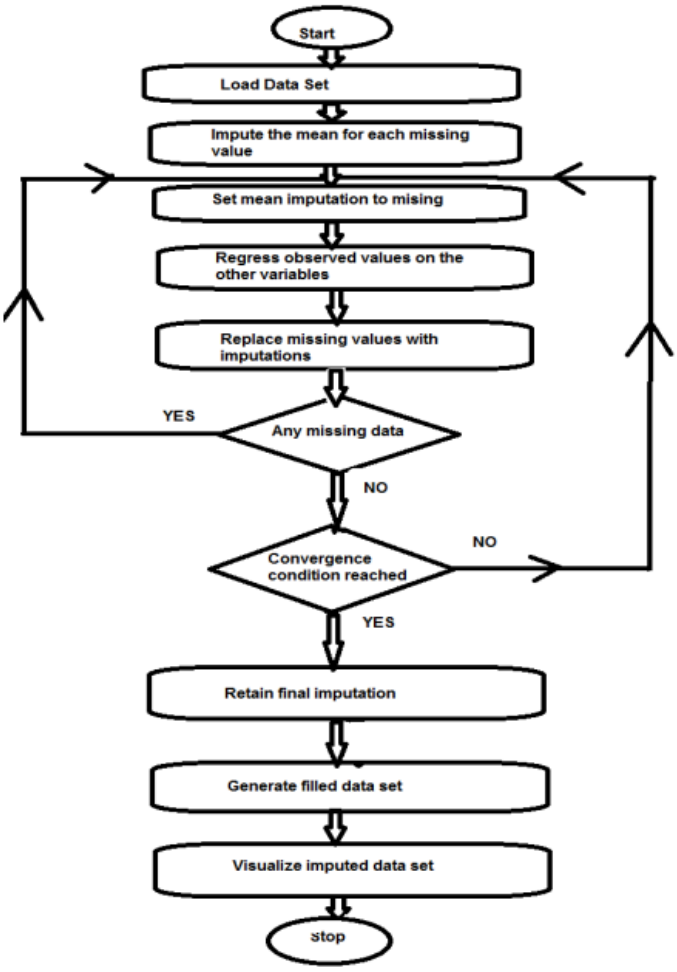
Osnovni princip MICE metode je da se nedostajuće vrednosti imputiraju jedna po jedna, koristeći druge dostupne atribute u skupu podataka. Ovaj proces se odvija kroz seriju koraka, gde se svaki atribut sa nedostajućim vrednostima tretira kao zavisan od drugih atributa. Na primer, ako imamo podatke o visini, težini i starosti, možemo koristiti visinu i starost da bismo procenili nedostajuću težinu.

Kada se nedostajuće vrednosti popune, MICE kreira više različitih setova podataka. Svaki od ovih setova ima svoje vlastite imputacije, što znači da se može proceniti koliko se vrednosti razlikuju. Ovaj pristup omogućava istraživačima da dobiju robusnije rezultate, jer uzimaju u obzir nesigurnost u vezi s nedostajućim vrednostima. Nakon što se podaci popune, rezultati iz različitih setova se kombinuju kako bi se dobile konačne procene.

Način na koji funkcioniše ova metoda se može razložiti na nekoliko koraka (Slika 10):

1. Kada se u skupu podataka jave nedostajući podaci, prva faza uključuje jednostavnu imputaciju, kao što je korišćenje srednje vrednosti. Ove srednje vrednosti mogu da služe kao privremene oznake za popunjavanje.

2. Ove privremene vrednosti zatim se ponovo dodeljuju mestima gde su podaci nedostajali.
3. U sledećem koraku, vrednosti koje su određene u prethodnoj fazi regresiraju se na ostale varijable iz modela. Ove druge varijable ne moraju obuhvatati sve atribute iz skupa podataka. U ovom sučaju, nedostajuća vrednost postaje zavisna varijabla unutar regresionog modela, dok ostale varijable deluju kao nezavisne.
4. Nakon regresije, nedostajuće vrednosti se zamenjuju predikcijama dobijenim iz modela. Kada se kasnije koristi ova varijabla kao nezavisna promenljiva u drugim modelima, koristiće se kako i stvarne i popunjene vrednosti.
5. Ovi koraci se ponavljaju za svaku varijablu koja ima nedostajuće vrednosti. Prolazak kroz sve varijable čini jednu iteraciju. Na kraju svake iteracije, nedostajuće vrednosti se zamene predikcijama koje odražavaju odnose među podacima.
6. Ovi procesi se nastavljaju kroz više iteracija, pri čemu se podaci ažuriraju posle svake iteracije. Moguće je postaviti broj iteracija koji treba sprovesti. Na kraju, konačne popunjene vrednosti se čuvaju, čime se dobija kompletan skup podataka.



Slika 10. MICE algoritam

4.2.5. Metode zasnovane na mašinksom učenju

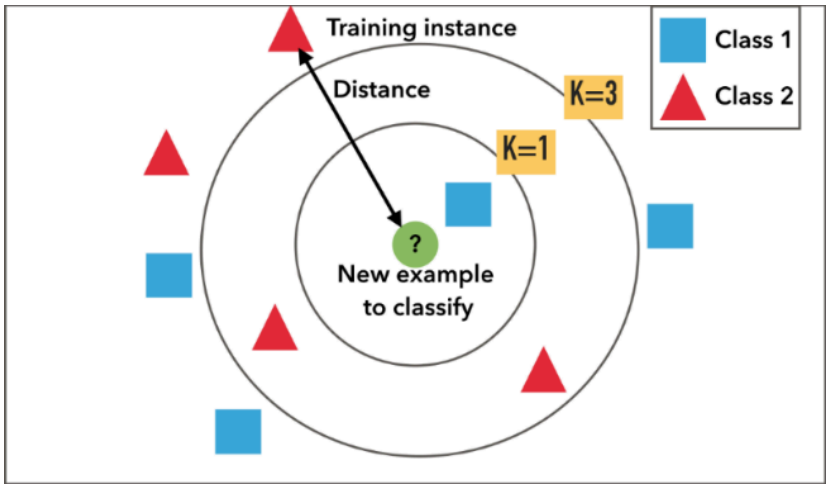
Metode imputacije koje se oslanjaju na mašinsko učenje koriste pristupe nadgledanog ili nenadgledanog učenja za procenu nedostajućih vrednosti u skupovima podataka. Ove metode se oslanjaju na dostupne informacije iz podataka koji nisu izgubljeni kako bi ostvarile preciznije predikcije. Metode mašinskog učenja mofu da prepoznaju složene veze i obrasce unutar podataka. Ove tehnike su fleksibilne, otporne na šum i izuzetke, a mogu se prilagoditi različitim tipovima podataka i obrascima nedostajućih informacija. Neki od uobičajenih pristupa uključuju regresiju, klasifikaciju i klasterizaciju. O regresiji je već bilo reči, te ćemo sada obraditi klasterizaciju.

4.2.5.1. KNN imputacija (Imputacija pomoću k najbližih suseda)

K-najbližih suseda je jednostavna metoda mašinskog učenja, gde je ideja da se nova tačka, čiji rezultat želimo da predvidimo, upoređuje sa drugim tačkama u skupu podataka koje već imamo.

Kada imamo novu tačku, KNN gleda koje su to K tačaka (suseda) najbliže njoj. Najbliže je najčešće definisano razdaljinom poput L2. Nakon što KNN pronađe K najbližih suseda, odluka se donosi na osnovu njihovih vrednosti (Slika 11). Ako rešavamo problem klasifikacije, nova tačka će dobiti klasu koja spada u većinu klasa njenih K suseda, a ako se radi o regresiji, prosto se uzima prosek vrednosti K tačaka da bi se predvidela nova vrednost.

Jedna od prednosti KNN-a je laka i jednostavna interpretabilnost i primena. Međutim, može biti sporiji kada radimo s velikim skupovima podataka, jer mora da uporedi novu tačku sa svim ostalim tačkama. Takođe, izbor vrednosti K može značajno uticati na rezultate.



Slika 11. KNN - Primer

5. Modeli mašinskog učenja otporni na nedostajuće podatke

Modeli mašinskog učenja koji su otporni na nedostajuće podatke mogu da rade čak i kada neki podaci nedostaju. Ovo je korisno jer smanjuje potrebu za popunjavanjem tih nedostajućih vrednosti pre nego što se analize urade.

Jedan od najpoznatijih pristupa su stabla odlučivanja (decision tree). Ovi modeli mogu da preskoče nedostajuće vrednosti dok donose odluke, što pomaže da rezultati budu tačniji.

Takođe, algoritmi poput Random Forest i XGBoost koriste tehnike koje im omogućavaju da uče iz podataka čak i kada nedostaju neki delovi. Ovi modeli prave više stabala odlučivanja i kombinuju njihove rezultate, što povećava tačnost.

5.1. Stabla odluke (Decision Tree) i Random forest

Stabla odluke su fleksibilni modeli mašinskog učenja koji se lako nose s nedostajućim podacima. Kada se gradi stablo, algoritam koristi atribute podataka da stvori čvorove koji dele podatke na manje grupe. Ove podele se zasnivaju na merenjima kao što su entropija, Gini indeks ili informacioni dobitak.

Jedna od glavnih prednosti stabala odluke je to što mogu uključiti nedostajuće vrednosti u proces podele bez potrebe za njihovim prethodnim popunjavanjem. Algoritam može da napravi posebne grane za podatke koji nedostaju, tretirajući ih kao poseban slučaj. Na primer, ako nedostaje informacija o nekoj karakteristici, algoritam može odlučiti kako će podeliti podatke koristeći druge dostupne karakteristike. Ova sposobnost omogućava modelima da budu otporni na nedostajuće podatke i da iskoriste sve informacije koje su dostupne prilikom izgradnje stabla.

Tehnike poput Random Forest, koje kombinuju više stabala odlučivanja, dodatno pomažu da se smanji uticaj nedostajućih podataka, jer svako stablo može da se nosi s njima na svoj način. Međutim, ako je stepen nedostajućih podataka vrlo visok ili ako postoje složeni obrasci nedostajanja, performanse stabla mogu biti smanjene. U tim slučajevima može biti korisno primeniti imputaciju nedostajućih vrednosti pre nego što se koristi stablo odlučivanja.

Metode zasnovane na stablima pokazuju dobru efikasnost u obradi podataka koji nedostaju nasumično (MCAR) ili su u zavisnosti od drugih varijabli (MAR), dok se u određenoj meri mogu nositi i sa podacima koji su izgubljeni na način koji nije nasumičan (MNAR). Pored toga, nedostajući podaci i izuzeci (outliers) imaju minimalan uticaj na algoritme stabla odluke.

6. Zaključak

Nedostajući podaci predstavljaju veliki problem u analizi informacija, jer mogu smanjiti tačnost i pouzdanost dobijenih rezultata. U ovom radu istražene su različite metode za popunjavanje nedostajućih vrednosti, koje se kreću od jednostavnih tehnika, poput zamene srednjom vrednošću, do složenijih pristupa, kao što je mašinsko učenje. Svaka od ovih metoda nosi svoje prednosti i slabosti, a izbor odgovarajuće tehnike često zavisi od specifičnih karakteristika skupa podataka i prirode nedostajućih vrednosti.

Upravljanje nedostajućim podacima zahteva pažljivo razmatranje i prilagođavanje izabranih metoda. Razumevanje različitih tehnika imputacije i njihovo pravilno korišćenje može značajno poboljšati kvalitet analize i omogućiti donošenje informisanih odluka i na osnovu nepotpunih informacija. Takođe, od velikog je značaja razumevanje samih podataka i pripadajućih atributa, kao i, ukoliko je to moguće, razumevanje povezanosti i potencijalne međusobne uslovljenosti atributa.

7. Literatura

Garret M. Fitzmaurice, Michael G. Kenward, Geert Molenberghs, Anastasios A. Tsiatis, Geert Verbeke. Handbook of Missing Data

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

Youran Zhou, Sunil Aryal. A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms