Универзитет у Нишу

Електронски факултет

Катедра за рачунарство

Прикупљање и предобрада података

Недостајући подаци

Ментор:

Проф. др. Александар Станимировић

Кандидат:

Стеван Грујић 1493

Ниш, 2023.

Садржај

1 Увод

1.1 Проблем недостајућих података

У савременим истраживањима, као и у свакодневном животу, подаци су кључни за доношење закључака и правилних одлука. Међутим, често се догађа да неки подаци нису доступни или недостају, што може бити последица различитих фактора, као што су грешке у прикупљању података, одбијање испитаника да одговоре на одређено питање или технички проблеми при снимању података. Ово стање се назива "недостајући подаци" („*missing data*“), и често може изазвати низ различитих проблема приликом саме анализе података.

Недостајући подаци могу утицати на тачност и поузданост анализе података, те довести до губитка информација, погрешних закључака, па чак и до непотпуне или погрешне интерпретације резултата. Због овога је важно разумети проблеме које недостајући подаци могу изазвати и применити адекватне методе за њихово руковање.

Један од кључних проблема је то што већина статистичких техника, како класичних тако и модерних, претпостављају (или захтевају) комплетне податке, а најчешћи статистички софтвери подразумевају најмање пожељне опције за руковање са недостајућим подацима, што може довести до брисања целог случаја из анализе. Највећи проблем је што ово може довести до избацивања важних података из анализе, упркос чињеници да би тај појединац или случај могао да допринесе укупној анализи са много других података. [1]

Да би се избегле потенцијалне последице недостајућих података, важно је разумети проблеме које они могу изазвати и применити адекватне методе за руковање са њима. У овом раду истражујемо проблем недостајућих података, анализирамо факторе који доводе до њиховог настанка и представљамо различите методе за руковање са њима. Такође, разматрамо предности и недостатке сваке методе како би се истакло шта треба узети у обзир приликом избора методе за руковање са недостајућим подацима у одређеном истраживању.

1.2 MCAR, MAR, MNAR

Иако би било веома пожељно утврдити разлог недостајања података у циљу прецизне анализе података, често се морамо задовољити само описима недостајућих података. Наиме, по завршетку прикупљања података, потребно је утврдити механизам по којем подаци недостају односно њихову дистрибуцију и зависност дистрибуције од измерених карактеристика узорка. Образац недостајања података много је важнији од саме количине недостајућих података. [2]

Ове обрасце, односно механизме недостајања података, можемо свести на три основна:

1. Потпуно случајно дистрибуирани недостајући подаци (енгл. *Missing Completely At Random - MCAR)*
2. Случајно дистрибуирани недостајући подаци (енгл. *Missing At Random - MAR)*
3. Подаци који не недостају по случајном распореду (енгл. *Missing Not At Random - MNAR*)

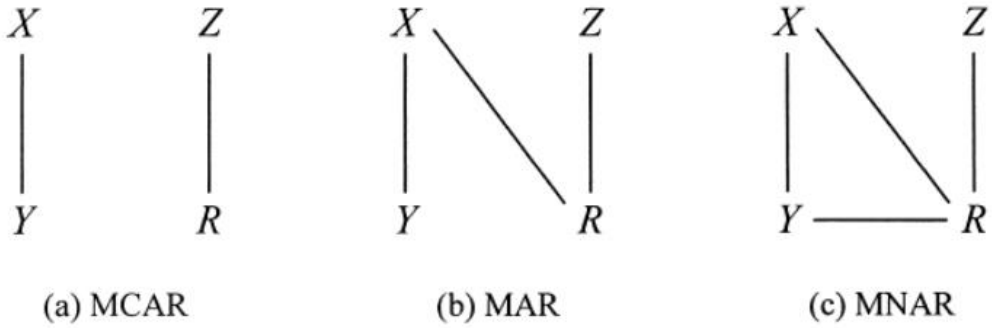
Недостајући подаци су типа ***MCAR*** ако је вероватноћа да податак недостаје независна од самих података. Другим речима, подаци су *MCAR* ако разлог за недостајуће вредности у излазу или предикторима нема везе са вредностима самих података, било да су постојећи или недостајући. Са друге стране, недостајући подаци су ***MAR*** ако је вероватноћа да податак недостаје независна од недостајућих вредности датих посматраних података. Другим речима, под *MAR*-ом можемо предвидети колико је вероватно да вредност недостаје на основу непотпуних података. На крају, недостајући подаци су ***MNAR*** ако, чак и уз вероватноћу да податак недостаје на основу свих посматраних информација, вероватноћа да податак недостаје зависи од самих незапажених недостајућих вредности.

Разлика између *MCAR* и *MAR* је у томе што изостајање података у оквиру *MCAR* није под утицајем ниједног атрибута у скупу података. С друге стране, изостајање података по *MAR* механизму није под утицајем атрибута у којем има недостајућих података, али јесте под утицајем неког другог атрибута у скупу података. Замислимо да нам је циљ истраживања да испитамо ставове према политици у односу на социо-демографске карактеристике. Врло је могуће, на пример, да испитаници с одређеним ставом не желе да дају одговор на питање о висини примања, и у том случају постоји пристрасност у давању одговора на питање о примањима (*MAR* механизам). Искључивањем узрока недостајућих података у описаном случају не бисмо добили непристрасне процене параметара у моделу. Укратко, не бисмо могли да генерализујемо резултате. Ипак, уколико подаци недостају по *MAR* механизму, недостајући подаци се могу објаснити расположивим подацима, јер су атрибути који су повезани са овим узроком измерени и могу се укључити у модел у циљу добијања непристрасних процена параметара.

Уколико, пак, дистрибуција недостајућих података зависи од самих недостајућих податка, кажемо да подаци не недостају по случајном распореду, односно да се дистрибуирају по *MNAR* обрасцу. У овом случају, недостајући податак је повезан с разлогом зашто недостаје. У складу са претходним примером, то би била тенденција испитаника с већим примањима да не извештавају о својим примањима. Анализе над оваквим подацима дају пристрасне процене параметара.

Како бисмо интуитивније схватили разлику између ових појмова, навешћемо један пример. Размотримо скуп података са следећа три атрибута: статус болести, ниво изложености и старост. Претпоставимо да за неке појединце недостаје изложеност (они имају вредност изложености, али не знамо шта је то). Шта сваки од три недостајућа механизма података подразумева у овој поставци?

1. *MCAR*: Било које две особе, без обзира на њихове вредности статуса болести, степена изложености и старости, имају исту вероватноћу да ће имати недостајућу вредност за изложеност.
2. *MAR*: Било које две особе са истим статусом болести и годинама имају исте шансе да имају недостајућу вредност изложености, без обзира на то колико је велики или мали њихов стварни ниво изложености.
3. *MNAR*: Чак и међу појединцима са истим статусом болести и годинама, шанса да им недостаје вредност изложености зависи од њиховог нивоа изложености.



Слика 1.2.1 Илустрација ова три механизма

На слици 1.2.1 илустрована су сва три механизма. X представља атрибут који садржи потпуно посматране податке (нема непостојећих података), Y представља атрибут који је недостајући, Z представља компоненту узрока нестанка која није у вези ни са једним од ових атрибута и R представља одсуство (*missingness*).

Недостајући подаци на нумеричким и категоричким атрибутима су уобичајени у истраживањима. Међутим, недостајући подаци на категоричким атрибутима могу представљати већи проблем, посебно када се ради о номиналним категоријама. Када се суочимо с оваквим проблемом, потребно је узети у обзир неколико чинилаца попут облика дистрибуције, количине недостајућих података, величине узорка, поузданости инструмената, и других релевантних фактора. Међутим, свако истраживање је јединствено, па стога, пре него што се крене у анализу, неопходно је разумети податке и узети у обзир све чиниоце који могу утицати на њихову поузданост и валидност. Без обзира на све чиниоце који утичу на обраду података, циљ постављен истраживањем треба да буде усмерен на опште препоруке које се могу применити у анализи и третману недостајућих података како би се постигли валидни и поуздани резултати.

2 Конвенционалне методе обраде недостајућих података

Постоје многи начини за обраду недостајућих података, од којих су неки традиционални, а неки модерни. Међу традиционалним методама су искључивање недостајућих података или једнострука импутација. Међу модерним методама су методе засноване на моделу, као што су методе максималне веродостојности и вишеструке импутације. Већина ових метода се може користити како за нумеричке, тако и за номиналне атрибуте. Међутим, неке од ових метода нису идеално решење за коришћење на номиналним атрибутима, док се друге препоручују.

2.1 Искључивање недостајућих података

Постоје два начина искључивања недостајућих података. Први начин се односи на **искључивање недостајућих података у целини** (енгл. *„Listwise deletion"*) тзв. паметно брисање са листе, а други подразумева тзв. **искључивање недостајућих података по паровима** (енгл. *„Pairwise deletion"*). Ови третмани недостајућих података су међу старијим методама, но њихово коришћење је још увек веома распрострањено.[2]

Искључивање података у целини, такође познато као анализа потпуних случајева, је статистичка метода која се користи у анализи података. Ова метода се користи како би се уклонили недостајући подаци из скупа података, тако да се узимају у обзир само случајеви који имају потпуне информације за све атрибуте које се проучавају. Дакле, уколико било који недостајући податак постоји у врсти, та врста се потпуно уклања из скупа података.

Ова метода се често користи због своје једноставности, јер не захтева додатне кораке како би се попунили недостајући подаци или развиле замене за њих. Међутим, *„Listwise deletion"* може имати велике недостатке, посебно ако постоји велики број недостајућих података. Уклањање великог броја редова може довести до смањења узорка и губитка значајних информација. Ово може довести до искривљења статистичких анализа и до неисправних закључака.

Грахам и сарадници наводе да су оба начина искључивања недостајућих података, генерално, неприхватљива. Томе у прилог иду резултати који показују да се искључивањем случајева губи на снази теста. Резултати у вези са искључивањем су прилично конзистентни – искључивање података у целини може дати непристрасне процене параметара уколико су недостајући подаци *MCAR* и ако недостаје мање од 5% случајева, али не и уколико су подаци *MAR* типа. Ипак, резултати наведених истраживања показују да третирање недостајућих података на овај начин резултира мање ефикасним проценама параметара, чак и кад су подаци *MCAR* типа. Иако ова процедура има очигледних мањкавости, у оквиру статистичких пакета често је аутоматски одабрана опција. Кинг и сарадници (Кинг, Хонакер и Јозеф, 2001) налазе да 94% анализа анкетних истраживања користе управо потпуно искључивање недостајућих података, и да аналитичари губе просечно трећину података на тај начин.

Искључивање по паровима подразумева искључивање из анализе само оних случајева који имају недостајуће податке на атрибутима на којима се врши анализа. На пример, уколико анализа почива на корелационој матрици, свака појединачна корелација из матрице ће бити рачуната посебно, на свим расположивим подацима. Иако се овим начином искоришћавају сви подаци којима располажемо, овај начин има још веће мањкавости у односу на претходно описани. Наиме, видимо и из претходног примера корелационе матрице, да ће свака корелација бити рачуната на различитом скупу података. Шта више, на овај начин могуће је чак генерисати интеркорелациону матрицу чија детерминанта није позитивно дефинитна. То би значило да матрица резултира негативним вредностима карактеристичних коренова, односно да је изолована варијанса неких компоненти негативна. Дакле, ова процедура може дати пристрасне процене параметара, управо због различитих узорака на којима су рачунате појединачне вредности (Грахам ет ал., 2003). Како не постоји јединствена величина узорка, не постоји ни основ за рачунање стандардних грешака параметара.

2.2 Једноструке импутационе методе

И *„Listwise deletion“* и *„Pairwise deletion“* генерално одбацују врсте података у којима се налазе и познати подаци (поред непостојећих). Атрактиван алтернативни приступ за руковање непотпуним подацима је импутација (попуна) вредности ставки које недостају. Могу се користити различити приступи импутацији који се крећу од изузетно једноставних до прилично сложених. Ове методе се могу применити за импутацију једне вредности за сваку ставку која недостаје (појединачна импутација) или, у неким ситуацијама, за импутацију више од једне вредности, да би се омогућила одговарајућа процена несигурности импутације (вишеструка импутација о којој ће бити речи мало касније).[3]

Једноструке импутационе методе су методе које се ослањају на једноставне статистичке методе како би се попунили недостајући подаци. Најчешће коришћене једноструке импутационе методе су замена недостајућих података средњом вредношћу, импутација коришћењем регресије и случајна импутација. Управо ћемо ове три методе описати у наставку.

2.2.1 Замена недостајућих података средњом вредношћу

Замена недостајућих података средњом вредношћу представља једну од метода која се користи за третирање недостајућих података у истраживањима. Идеја је да се, у одсуству других информација, средња вредност узорка или групе користи за сваког појединачног учесника који има недостајући податак. У случају нумеричких атрибута, овај третман подразумева да се недостајући податак замени аритметичком средином добијеном на случајевима који имају податак на датом атрибуту. У случају номиналних атрибута, замена би се морала вршити модалном вредношћу, а у случају

ординалних – медијаном. Међутим, ова метода није увек најбољи избор јер може да доведе до кривљења података и погрешних закључака.

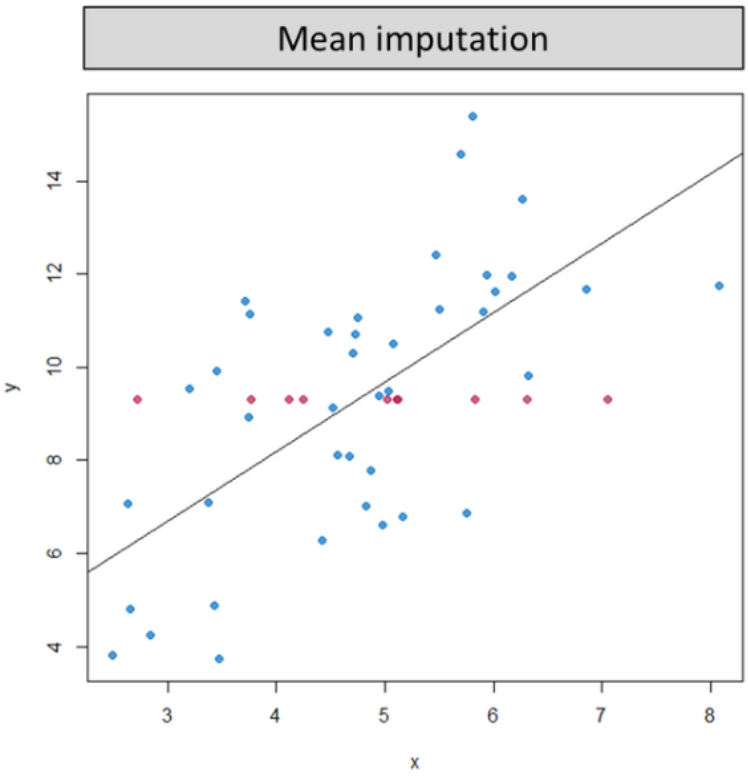
Један од проблема са овом методом је да она може да смањи варијансу података, чак и ако недостајући подаци нису повезани са другим атрибутима. На пример, ако 20% узорка има недостајуће податке, а средња вредност се користи за замену ових података, то ће смањити варијансу променљиве и тиме довести до погрешних закључака. Што је већи проценат недостајућих података, то ће ефекат бити израженији.

Ова метода такође може да доведе до лоше процене стандардне девијације података, што значи да се процена варијансе података неће тачно поклопити са стварном варијансом у популацији. Чак и у случајевима када су подаци недостајући случајно (*MCAR*), употреба средње вредности за замену недостајућих података може да доведе до искривљења резултата.

Ово можемо илустровати следећим примером: Посматрајмо скуп података о математичким успесима ученика, који укључује информације о годинама, полу и оценама. Међутим, у неким случајевима немамо информације о оценама, јер су неки ученици били одсутни током тестова или су пропустили да упишу своје резултате.

Ако бисмо уместо недостајућих оцена користили средњу вредност за све ученике који нису предали своје резултате, то би могло да доведе до искривљења резултата. Наиме, уколико је велики број ученика предало своје резултате, али је неколико њих то пропустило да уради, то значи да ће се средња вредност изузетно мало разликовати од просечне оцене коју би имали када би сви ученици предали резултате. Ово ће утицати на то да се чини да је разлика између најбољег и најлошијег ученика мања него што је то заиста случај, јер нећемо имати информације о доприносу ученика који нису предали резултате у укупним резултатима.

У неким случајевима, брисање недостајућих података може да буде боља опција јер се тиме задржава оригинална варијанса података и тиме се добијају прецизнији резултати. Укратко, замена недостајућих података средњом вредношћу није увек најбоља опција за третирање недостајућих података у истраживањима. Она може да доведе до искривљења података и погрешних закључака, поготово када је проценат недостајућих података висок. Због тога, препоручује се коришћење других метода које могу да обезбеде тачније резултате.



Слика 2.2.1 Илустрација замене средњом вредношћу

2.2.2 Случајна импутација

Hot deck импутација недостајућих вредности је једна од најједноставнијих метода за једноструку импутацију. Метода је интуитивно очигледна, а главна идеја је да случај са недостајућом вредношћу добија валидну вредност која се узима насумично из случајева који су највише слични оном који недостаје, на основу неких позадинских атрибута које одреди корисник (ови атрибути се називају "*deck* атрибути"). Скуп донаторских случајева назива се " *deck* ". [4]

У најосновнијем сценарију, када имамо скуп података који садржи само један атрибут (садржи недостајуће вредности), можемо за *deck* атрибут изабрати управо поменути једини атрибут (На тај начин овај атрибут постаје тзв. „Позадински атрибут“), након чега се врши случајни избор од *n-m* валидних случајева да буду донатори за *m* случајева са недостајућим вредностима. Случајна замена је суштина *hot deck* методе.

Да би се омогућила идеја о корелисаности која утиче на вредности, користи се усклађивање са специфичнијим позадинским атрибутима (атрибутима који су на неки начин повезани са циљаним атрибутом који садржи непостојеће податке). На пример, можемо да импутирамо недостајући одговор белог мушкарца у доби од 30-35 година од донатора који припадају тој специфичној комбинацији карактеристика

(бели мушкарац, 30-35 година). Позадинске карактеристике би требало, барем теоријски гледано да буду повезане са анализираном карактеристиком (за коју се врши импутација).

Hot-deck импутација је стара, али и даље популарна јер је и једноставна по идеји и, у исто време, погодна за ситуације у којима методе обраде недостајућих вредности као што су искључивање недостајућих података у целини или замена недостајућих података средњом вредношћу неће бити од користи јер недостаци који се налазе у подацима нпр. нису типа *MCAR,* већ типа *MAR. Hot-deck* је прилично погодан за *MAR* шаблон с обзиром на то да је у овом механизму недостајање података изазвано неким познатим атрибутом (атрибутом који се налази у скупу података и који нема недостајућих вредности), а што је врло слично начину на који ради *hot-deck.*

Главни недостатак *hot-deck* импутације је тај што захтева да горепоменути позадински атрибути буду свакако номинални (због номиналности, није потребан никакав посебан „алгоритам подударања“). Уколико нема таквих, потребно је да континуалне атрибуте дискретизујемо и на тај начин направимо категорије. Што се тиче променљивих са недостајућим вредностима - оне могу бити било које врсте, што је свакако једна од главних предности ове методе (многи алтернативни облици појединачне импутације могу да импутирају само континуалне атрибуте).

2.2.3 Једнострука импутација помоћу регресије

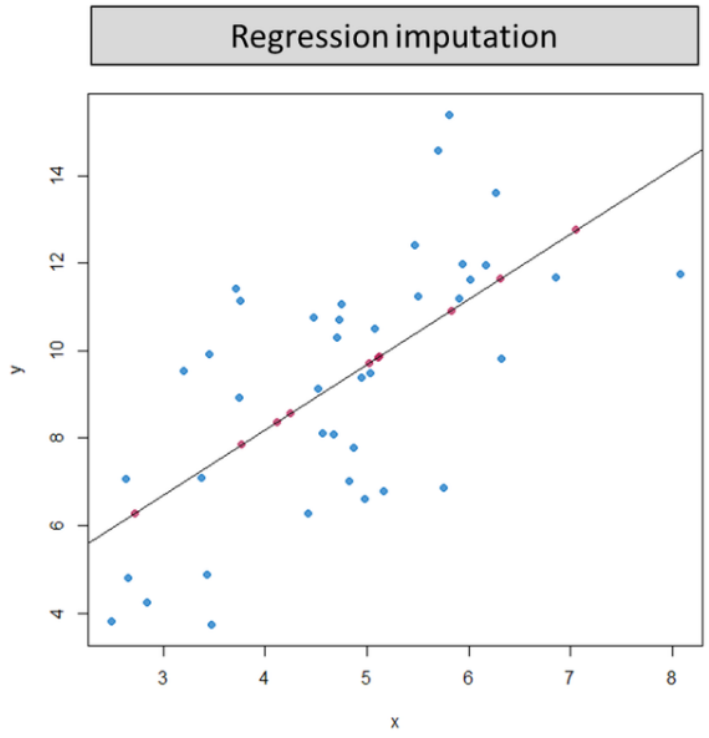
Једнострука импутација помоћу регресије (енг. *Single imputation via regression*) је метода замене недостајућих података коришћењем линеарне регресије. Ова метода се користи када постоји зависност (висока корелација) између недостајућих података и осталих података у скупу података.

Основна идеја једноструке импутације помоћу регресије је да се користи линеарна регресија између недостајућег податка и осталих података у скупу података како би се предвидела вредност недостајућег податка. Конкретно, линеарна регресија се примењује на подацима који су комплетни како би се одредили коефицијенти једначине регресије. Затим се ови коефицијенти користе за предвиђање вредности недостајућег податка за сваку јединствену вредност у скупу података.

Као што је већ напоменуто, ова метода може бити врло корисна када постоји јак линеарни однос између недостајућег податка и осталих података у скупу података. Међутим, када постоји слаба или нема никакве везе између недостајућег податка и осталих података, ова метода може довести до погрешних закључака и искривљених резултата, као што је то раније био случај са заменом недостајућих података средњим вредностима. Такође, ова метода може погоршати проблем недостајућих података ако се користи на подацима који су вишеструко недостајући.

На пример, замислимо скуп података који садржи информације о старости, полу, висини и тежини особа. Ако у овом скупу података недостаје неколико вредности тежине, можемо применити једноструку импутацију помоћу регресије како бисмо предвидели недостајуће вредности тежине користећи старост, пол и висину као предикторе. У овом случају, подаци о старости, полу и висини представљају предикторе, а тежина је циљна променљива. Регресијска једначина се формира из података о комплетним подацима, а затим се примењује на недостајуће вредности тежине како би се предвиделе њихове вредности.

Међутим, ова метода има неколико ограничења. Као прво, једнострука импутација помоћу регресије може довести до прецењивања тачности модела и лоше процене варијабилности података. Такође, ова метода може довести до кривљења веза између различитих атрибута у подацима, што може довести до нетачних закључака.



Слика 2.2.2 Илустрација регресионе импутације

2.2.4 Једнострука импутација помоћу К најближих суседа (*KNN)*

Најближи сусед (*NN*) је метод који користи алгоритам надгледаног учења. Надгледано учење има за циљ проналажење нових образаца у подацима повезивањем постојећих образаца података са новим подацима. Постоје две врсте ових алгоритама, 1*NN* и *KNN*. Најближи сусед (1*NN*) је приступ који врши

класификацију у зависности од једног податка који је најближи посматраном недостајућем податку, док је *KNN* приступ који врши класификацију у зависности од К најближих података, при чему је К > 1.

KNN је метода која се користи за класификацију објеката на основу неких података који су најближи објекту. У класификацији, *KNN* ради тако што израчунава растојање између тест скупа података и података чија је класа позната (тренинг скупа) користећи, на пример Еуклидову удаљеност. Руковање подацима који недостају помоћу *KNN*-а почиње одређивањем броја најближих суседа или најближих опсервација које симболизује К, а затим израчунавањем најмање удаљености од сваког посматрања које не садржи податке који недостају. Кораци за импутирање података који недостају помоћу *KNN* методе су следећи:

1. Израчунавање растојања: Први корак у *KNN* импутацији је израчунавање растојања између недостајућих података и свих осталих података у скупу података. Постоји више начина за израчунавање растојања, а неки од најчешће коришћених метода су Еуклидско растојање, косинусно растојање и Махаланобисово растојање. Следи приказ формуле за рачунање Еуклидског растојања:

$$d(x_a x_b) = \sqrt{\sum_{j=1}^{m}(x_{aj} - x_{bj})^2}$$

2. Сортирање растојања на основу посматрања у опадајућем редоследу.

3. Одабир К суседа: Након израчунавања растојања, следећи корак је одабир К суседа са најмањим растојањем до недостајућег податка. К представља број суседа који ће бити узети у обзир, а обично се одабире вредност између 3 и 10.

4. Рачунање просечне вредности: Након одабира К суседа, следећи корак је рачунање просечне вредности недостајућег податка. Просечна вредност се рачуна као аритметичка средина вредности К суседа. У случају нумеричких података, просечна вредност се рачуна једноставно као средња вредност, док се за категоричке податке користи мода.

$$\overline{x_j} = \frac{\sum_{k=1}^{K} wk\ vk}{\sum_{k=1}^{K} wk},$$

где је vk вредност допуњеног податка, а wk је тежина К − те најближе обзервације.

5. Попуњавање недостајућих вредности: Коначни корак у *KNN* импутацији је попуњавање недостајућих вредности у скупу података. Недостајуће вредности се попуњавају просечном вредношћу која је израчуната у претходном кораку.

6. Евалуација резултата: Након импутације, потребно је евалуирати резултате и утврдити да ли је квалитет импутације задовољавајући. Ово се може урадити поређењем импутираних података са оригиналним подацима, или применом

неке друге методе за импутацију која ће служити као референца. Такође је важно узети у обзир да *KNN* импутација може бити осетљива на присуство екстремних вредности и треба водити рачуна о овом фактору при евалуацији резултата.

Једна од предности ове методе је што је прилагодљива и може да ради са различитим врстама података, укључујући нумеричке, категоричке и бинарне податке. Осим тога, *KNN* импутација је релативно једноставна за имплементацију и може да буде ефикасна у случајевима када се недостајуће вредности јављају у малом броју.

Ипак, *KNN* импутација може да има неке недостатке. На пример, ова метода не узима у обзир зависности између различитих атрибута у скупу података, што може да доведе до неадекватне импутације. Такође, избор параметра К, односно броја суседа који се узимају у обзир, може да утиче на квалитет импутације и потребно је правилно одабрати овај параметар.

Узимајући у обзир предности и недостатке *KNN* импутације, ова метода може бити корисна у различитим ситуацијама када је потребно попунити недостајуће податке у скупу података. Међутим, препоручује се пажљиво размотрити предности и недостатке ове методе и упоредити је са другим методама импутације пре него што се одлучите за њу.

3 Вишеструка импутација (*Multiple Imputation*)

Вишеструка импутација (*MI*) формално је представљена од стране Рубина (1978). Рубин пружа свеобухватно обрађивање ове технике. Кључна идеја поступка вишеструке импутације је замена сваке недостајуће вредности скупом од М вероватних вредности, тј. вредности "извучених" из дистрибуције података, које представљају потенцијалне податке у вези са правом вредношћу за импутирање. Импутирани скупови података се затим анализирају применом стандардних процедура за комплетне податке, након чега се комбинују резултати ових анализа.

Вишеструка импутација, барем у основној форми, захтева да механизам недостајућих података буде *MAR*. Међутим, техника је примењена и у *MNAR* окружењима. Вишеструка импутација укључује три различите фазе или, користећи Рубинову (1987) терминологију, задатке:

1. Недостајуће вредности се попуњавају М пута да би се генерисало М комплетних скупова података.
2. М комплетних скупова података се анализирају применом стандардних процедура.
3. Резултати М анализа се комбинују у једно закључивање.

Multiple Imputation (*MI*) има бројне предности у односу на друге приступе обради недостајућих података. Вишеструка импутација укључује попуњавање недостајућих вредности више пута, стварајући више "потпуних" скупова података. Описано детаљно од стране Schafera и Грахама (2002), недостајуће вредности се импутирају на основу посматраних вредности за одређену особу и односа посматраних атрибута у подацима за друге учеснике, претпостављајући да су посматрани атрибути укључени у модел импутације. Поступци вишеструке импутације, посебно *MICE (Multivariate Imputation by Chained Equations)*, су врло флексибилни и могу се користити у широком спектру поставки. Будући да вишеструка импутација укључује стварање више предвиђања за сваку недостајућу вредност, анализе вишеструко импутираних података узимају у обзир неизвесност у импутацијама и дају тачне стандардне грешке. На једноставном нивоу, ако у посматраним подацима (коришћеним у моделу импутације) нема много информација о недостајућим вредностима, импутације ће бити веома варијабилне, што доводи до високих стандардних грешака у анализама. Насупрот томе, ако посматрани подаци у великој мери предвиђају недостајуће вредности, импутације ће бити доследније између импутација, што доводи до мањих, али ипак тачних, стандардних грешака.[5]

3.1 MICE (Multivariate Imputation by Chained Equations)

MICE (*Multivariate Imputation by Chained Equations*) је посебна техника вишеструке импутације. *MICE* функционише на претпоставци да су подаци који недостају *Missing At Random* (*MAR*) с обзиром на атрибуте које се користе у поступку импутације, што значи да вероватноћа да вредност недостаје зависи само од посматраних вредности, а не од непосматраних вредности (описано у првом поглављу овог рада). Другим речима, након контроле свих доступних података (тј. атрибута које су укључене у модел импутације) "сва преостала одсутност података је потпуно случајна". Примена *MICE*-а када подаци нису *MAR* могла би резултирати пристрасним проценама. У остатку овог текста претпостављамо да се *MICE* поступци користе за податке који су *MAR*.

Многе првобитно развијене процедуре вишеструке импутације претпостављале су велики заједнички модел за све атрибуте, попут заједничке нормалне дистрибуције. У великим скуповима података са стотинама атрибута различитих врста, то је ретко прикладно. *MICE* је алтернативни, флексибилни приступ овим заједничким моделима. Заправо, *MICE* приступи су коришћени у скуповима података са стотинама и хиљадама атрибута. У поступку *MICE*-а, серија регресионих модела се покреће при чему се свака варијабла (атрибут) са недостајућим подацима моделује условно према другим атрибутима у подацима. Ово значи да се сваки атрибут може моделовати у складу са својом дистрибуцијом, нпр. Бинарни атрибути моделовани су помоћу логистичке регресије, а континуирани атрибути моделовани су помоћу линеарне регресије.

Процес ланчане једначине може се разложити на четири општа корака:

1. Корак 1: За сваки податак који недостаје у скупу података, врши се једноставна импутација, као што је импутација са средњом вредношћу. Ове средње испуне могу се сматрати „индексима места“.

2. Корак 2: „Индекси места“ импутације средње вредности једне варијабле („вар“) се враћају на вредности које недостају.

3. Корак 3: Уочене вредности променљиве „вар“ у кораку 2 регресирају се на друге варијабле у моделу популације, које могу, али не морају укључивати сви атрибути у скупу података. Другим речима, „вар“ је зависни атрибут у регресионом моделу, а све остале варијабле су независне у регресионом моделу. Ови модели регресије функционишу под истим претпоставкама које бисте направили када користите линеарне или логистичке регресионе моделе ван контекста попуњавања података који недостају.

4. Корак 4: Недостајуће вредности за „вар“ се затим замењују предвиђањима (испунама) из регресионог модела. Када се "вар" касније користи као независна променљива у регресионим моделима за друге варијабле, користиће се и посматране вредности и ове попуњене вредности.

5. Корак 5: Кораци 2–4 се понављају за сваку променљиву којој недостају подаци. Пролазак кроз сваку променљиву чини једну итерацију или "циклус". На крају једног покретања, све недостајуће вредности су замењене предвиђањима из регресија које су одражавале односе забележене у подацима.

6. Корак 6: Кораци 2-4 се понављају у неколико циклуса, а комплементи се ажурирају након сваког циклуса. Истраживач може одредити број циклуса које треба извести. Након ових циклуса, коначни комплементи се задржавају, што резултира једним подстављеним скупом података. Генерално се изводи 10 циклуса, међутим потребно је истраживање да би се одредио оптималан број циклуса за допуну података под различитим условима. Идеја је да на крају циклуса дистрибуција параметара који регулишу допуну (нпр. коефицијенти у регресионим моделима) конвергирају у смислу постизања стабилности. Ово ће, на пример, избећи зависност од редоследа променљивих допуна. У пракси, истраживачи могу да верификују конвергенцију упоређивањем регресионих модела у наредним серијама. Различити *MICE* софтверски пакети се разликују по тачној имплементацији овог алгоритма (нпр. по редоследу попуњавања променљивих), али општа стратегија је иста.

Како бисмо приближили приступ "везаних једначина", замислимо једноставан пример у којем имамо три атрибута у нашем скупу података: старост, искуство и приход, а све три имају барем неке недостајуће вредности (Слика 3.1.1).

age	experience	salary
25		50
27	3	
29	5	110
31	7	140
33	9	170
	11	200

Слика 3.1.1 Иницијални скуп података

MAR претпоставка би имплицирала да вероватноћа да одређени атрибут недостаје зависи само од посматраних вредности, и да, на пример, то да ли нечији приход недостаје не зависи од њиховог (непосматраног) прихода. У кораку 1 процеса *MICE*, сваки атрибут би прво био допуњен, на пример, просечном вредношћу, привремено постављајући било коју недостајућу вредност једнаку просечној посматраној вредности за тај атрибут (Слика 3.1.2).

age	experience	salary
25	7	50
27	3	134
29	5	110
31	7	140
33	9	170
29	11	200

Слика 3.1.2 Додате средње вредности

Затим, у кораку 2, допуњене просечне вредности старости би се поново поставиле на недостајуће вредности (Слика 3.1.3).

age	experience	salary
25	7	50
27	3	134
29	5	110
31	7	140
33	9	170
	11	200

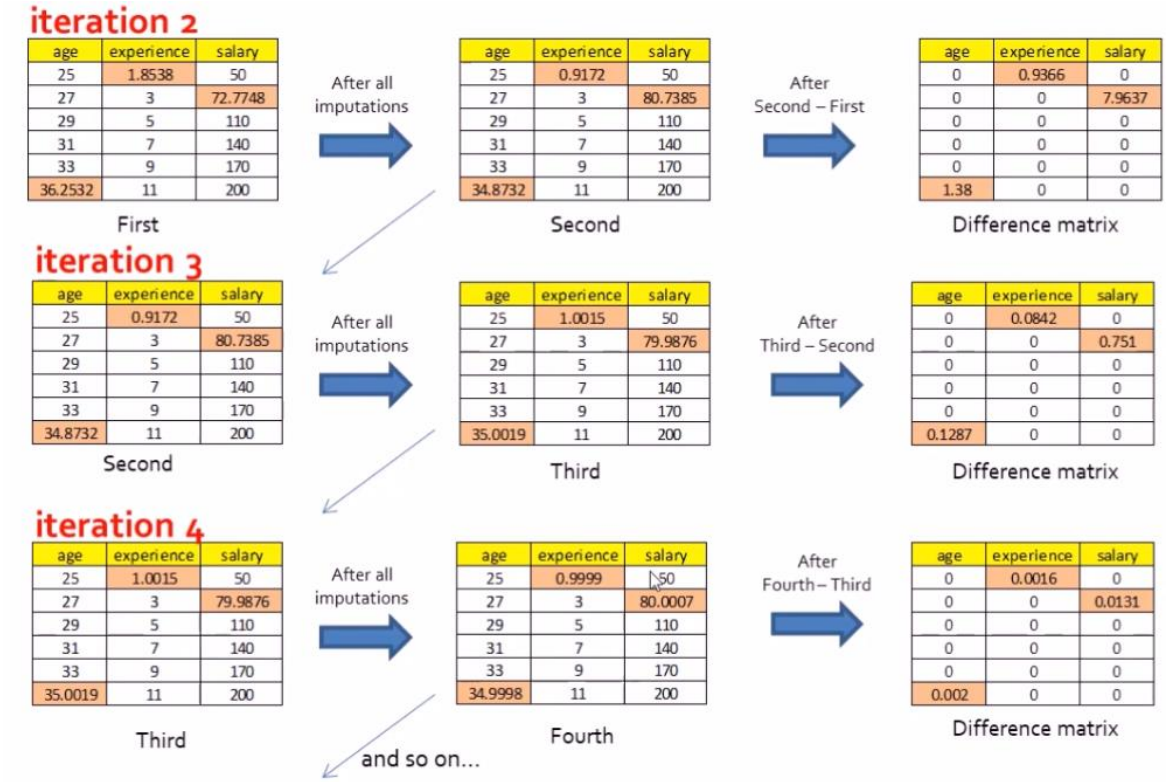
Слика 3.1.3 Поновно враћање на недостајуће вредности

У кораку 3, линеарна регресија старости предвиђена искуством и приходом би се покренула користећи све случајеве у којима је старост посматрана. У кораку 4, предвиђања недостајућих вредности старости добијена би из те једначине регресије и допуњена. У овом тренутку, старост више нема недостајућих вредности (Слика 3.1.4).

age	experience	salary
25		50
27	3	134
29	5	110
31	7	140
33	9	170
36.2532	11	200

Слика 3.1.4 Рачунање коришћењем линеарне регресије

Кораци 2–4 би се потом понављали за атрибут искуства. Оригинално недостајуће вредности искуства би се поново поставиле на недостајуће, а линеарна регресија прихода предвиђена старошћу и приходом би се покренула користећи све случајеве са посматраним искуством. Предвиђања (предвиђене вредности) би се добијала из те једначине регресије за недостајуће вредности прихода. Затим би се кораци 2–4 поново поновили за атрибут прихода. Оригинално недостајуће вредности прихода би се поставиле поново на недостајуће и линеарна регресија прихода према старости и искуству би се покренула користећи све случајеве са посматраним приходом. Предвиђања из тог модела линеарне регресије би се користила за допуну недостајућих вредности прихода. Цео овај процес итерирања кроз три атрибута би се понављао све док се не постигне конвергенција; посматрани подаци и коначни скуп допуњених вредности би онда чинили један "потпун" скуп података (Слика 3.1.5).



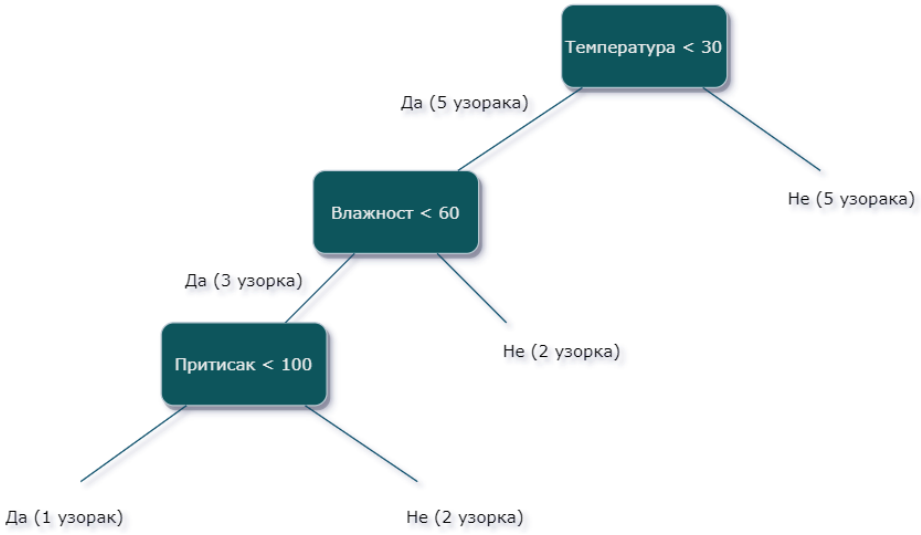
Слика 3.1.5 Конвергенција ка нули

MICE је популаран алгоритам импутације недостајућих података који се често користи и у пракси. Овај алгоритам има многе предности, као што су флексибилност, могућност рада са различитим типовима података, као и способност да узме у обзир корелацију међу атрибутима.

Укратко, *MICE* алгоритам је ефикасан начин за решавање проблема недостајућих података, што је често проблем у реалним подацима. Међутим, као и код сваког алгоритма, важно је узети у обзир ограничења и могуће недостатке, а такође треба бити опрезан приликом интерпретације резултата који су добијени након примене овог алгоритма.

4 Модели машинског учења отпорни на недостајуће податке

Многи популарни модели машинског учења, као што су машине за подршку векторима (*Support Vector Machines*), *glmnet*, неуронске мреже, итд, не могу толерисати било какву количину недостајућих података. Међутим, постоји неколико предиктивних модела који могу интерно да обрађују непотпуне податке. Одређене имплементације модела заснованих на стаблу имају паметне процедуре за прилагођавање непотпуних података. *CART (Classification And Regression Tree)* методологија користи идеју сурогатних подела (*surrogate split*). Приликом креирања стабла, формира се засебан скуп поделе (коришћењем других предиктора од тренутног предиктора који се дели) који могу апроксимирати изворну логику поделе уколико недостаје вредност за тај предиктор. Слика 4.1.1 приказује модел рекурзивног партиционисања. [6]



Слика 4.1.1 Стабло одлуке

У овом стаблу одлучивања коришћена су три предиктора: температура, влажност и притисак. Међутим, неким узорцима недостаје вредност за температуру. У том случају, за доношење одлуке може се користити сурогат предиктор, на пример, влажност. Ако је вредност влажности испод 60%, одлука се доноси на основу притиска, ако је притисак испод 100, узорак класификујемо као „Да“, а ако није, класификујемо га као „Не“. На овај начин, стабло одлучивања које је типа *CART* може да се прилагоди подацима који недостају и да и даље буде корисно за класификацију нових узорака. Наравно, сходно овоме, *Random Forest* алгоритам је такође отпоран на недостајуће податке.

Поред наведеног алгоритма, треба споменути и *Naïve Bayes* алгоритам. Наивни Бајесов алгоритам је пробабилистички алгоритам за класификацију који, такође, може толерисати недостајуће податке. У наивном Бајесовом алгоритму, вероватноћа да узорак припада некој класи се израчунава на основу условних вероватноћа

атрибута за ту класу. У случају недостајућих вредности за неки атрибут, наивни Бајесов алгоритам једноставно изоставља тај атрибут приликом израчунавања вероватноће.

На пример, ако имамо скуп података који се састоји од три атрибута - старост, пол и приход, при чему постоји недостајућа вредност за приход за неке узорке, наивни Бајесов алгоритам ће једноставно изоставити тај атрибут за те узорке. Када израчунава вероватноћу да узорак припада некој класи, алгоритам ће узети у обзир само старост и пол као атрибуте за те узорке. Битно је напоменути да, у случају овог алгоритма, с обзиром на поменути начин рада, може доћи до губитка битних информација из скупа података.

Укратко, коришћење модела који су отпорни на недостајуће податке је од кључног значаја у многим областима машинског учења. Модели као што су Наивни Бајес и *Random Forest* су показали високу ефикасност у раду са подацима који садрже недостајуће вредности. Међутим, важно је напоменути да, иако су ови модели отпорни на недостајуће податке, може се, итекако, доћи до губитка информација што може утицати на квалитет предикција. Стога, потребно је пажљиво приступити одабиру модела и проценити ризике и предности њиховог коришћења у зависности од врсте података са којима радимо.

Како би се постигла најбоља могућа ефикасност у раду са недостајућим подацима, свакако је најбоља пракса користити неке од метода које су описане раније у овом раду. Све ове технике имају своје предности и мане и важно је пажљиво одабрати најбољи приступ у складу са природом података и циљевима које желимо постићи.

5 Практични део рада

У практичном делу овог рада, истраживали смо различите алгоритме за обраду недостајућих података (алгоритме описане у теоријском делу рада) и њихов утицај на перформансе различитих модела машинског учења. Коришћен је скуп података о квалитету воде за пиће, који је доступан на Каггле платформи. Састоји се од информација о хемијским карактеристикама воде, као и информација о томе да ли је вода погодна за људску потрошњу или не. У скупу података се налази укупно 3276 инстанци, а за сваку инстанцу су дате информације о 9 различитих карактеристика воде, као и информација о томе да ли је вода погодна за људску потрошњу или не. Карактеристике воде су дате у нумеричком облику, а између њих постоје значајне разлике у распону вредности. Скуп података је интересантан за анализу јер се бави питањем квалитета воде које има важне импликације на здравље људи и животну средину.

Циљ овог истраживања био је да се упореде перформансе различитих алгоритама за обраду недостајућих података, као и да се одреди који модел машинског учења даје најбоље резултате за овај скуп података. У складу са тим, тестирали смо неколико

алгоритама за обраду недостајућих података, укључујући *MICE* импутацију, *KNN* импутацију, замену недостајућих података средњом вредношћу и медијаном, као и брисање недостајућих вредности.

Наши резултати показали су да је брисање недостајућих вредности дало најбоље перформансе у односу на остале методе обраде недостајућих података. Ова изненађујућа чињеница може се објаснити једино величином скупа података. С обзиром на то да скуп података није велики, ово има смисла. Осим тога, *Random Forest* се показао као најбољи модел машинског учења за овај скуп података, са комбинованом Ф-мером од 0,539 и подручјем испод *ROC* криве од 0,647.

Ова студија такође је показала да једнострука регресија даје сличне резултате као и *MICE* импутација, што указује на то да се ове две методе могу користити као алтернатива једна другој у ситуацијама када је потребно извршити обраду недостајућих података. Наравно, све ово искључиво зависи од саме природе података. Ако је то био случај са овим скупом података, не мора нужно да значи да ће таква ситуација бити са неким другим скупом.

Међутим, *KNN* импутација, импутација средњом вредношћу и медијаном се нису показале као добре методе за обраду недостајућих података у овом случају. *KNN* импутација је дала нешто боље резултате од импутација средњом вредношћу и медијаном, али је и даље далеко од перформанси добијених брисањем недостајућих вредности.

На крају, ова студија има важну практичну примену, јер показује да брисање недостајућих вредности може бити добра метода за обраду недостајућих података, али да треба бити јако обазрив са њом јер може доћи до губитка велике количине информација. Такође, ово истраживање указује на то да *Random Forest* може бити добар избор модела машинског учења.

6 Закључак

У овом семинарском раду упознали смо се са проблемом недостајућих података и различитим врстама података који недостају – *MCAR*, *MAR* и *MNAR*. Такође смо размотрили различите конвенционалне методе за руковање подацима који недостају, као што су искључивање података који недостају, методе појединачне импутације и вишеструка импутација.

Искључивање података који недостају је, као што смо видели, једноставан, али веома неефикасан метод. Методе појединачне импутације, као што је замена недостајућих података средњом вредношћу, случајна импутација, појединачна импутација помоћу регресије и К најближих суседа (*KNN*), су лаке за примену, али могу дати нетачне резултате и имати ограничења у погледу обраде сложенијих података.

С друге стране, вишеструка импутација нам омогућава да креирамо више верзија података које садрже различите процене вредности које недостају. Метода вишеструке импутације се најчешће користи када постоје значајне количине података који недостају, а предност има узимање у обзир несигурности око процене вредности које недостају.

Једна од најпопуларнијих техника вишеструке импутације је *MICE* (Мултиваријантна импутација помоћу ланчаних једначина), која се може користити за обраду сложенијих података. *MICE* користи серију појединачних модела импутације и користи се заједно са анализом која укључује варијансе између различитих процена вредности које недостају.

Поред овога, обрадили смо и неколицину модела који нативно подржавају недостајуће податке и као такви дају релативно добре резултате. Међутим, дошли смо до закључка да се горенаведене методе импутације ипак боље показују у пракси од ових модела, што итекако има смисла.

Узимајући у обзир наведено, могуће је закључити да постоји много начина за обраду података који недостају, а избор методе зависи од много фактора, укључујући врсту података, количину и распоред података који недостају, циљни атрибут и личне преференције истраживача. Неки истраживачи преферирају искључивање података који недостају како би избегли могуће непрецизне процене, док други преферирају вишеструку импутацију како би укључили несигурност у проценама вредности које недостају.

Важно је нагласити да ниједан метод за обраду података који недостају није савршен, а избор методе зависи од карактеристика података и циљева истраживања. Такође, треба бити опрезан у примени метода за руковање подацима који недостају, јер неке методе могу довести до искривљених резултата или повећане варијансе. У табели 6.1 можемо видети предности и мане описаних алгоритама, али и препоруку када је погодно користити их.

Алгоритам за обраду недостајућих података	Добре стране	Лоше стране	Препоруке за коришћење
Искључивање недостајућих података	Једноставан и брз	Може довести до губитка значајних података	Препоручује се само ако је проценат недостајућих података мали
Замена недостајућих података средњом вредношћу	Једноставан и брз, одржава статистичке карактеристике података	Може да измени дистрибуцију података	Овај приступ се препоручује ако је број недостајућих података релативно мали, а њихов утицај на резултате анализе није значајан.
Случајна импутација	Одржава статистичке карактеристике података, може да смањи шум у подацима	Може да изазове превелику корелацију са суседним подацима	Препоручује се само ако су недостајуће вредности (*MCAR)*
Једнострука импутација помоћу регресије	Може да одржава статистичке карактеристике података и да уклони шум. Узимајући у обзир друге атрибуте	Импутиране вредности су често превише прецизне и доводе до прецењивања корелације са другим атрибутима	Препоручује се коришћење само ако постоји јака веза између недостајуће вредности и других атрибута
Једнострука импутација помоћу *KNN*	Једноставан за примену; Ефикасан за мале и средње скупове података	Може бити веома непоуздан за велике скупове података; Не узима у обзир корелацију између атрибута	Препоручује се за мале и средње скупове података који имају ниске стопе недостајућих података
MICE алгоритам	Узима у обзир корелацију између атрибута; Може се применити над различитим типовима података; Може се применити на великим скуповима података	Спор за велике скупове података; Захтева више корака за обраду;	Препоручује се за велике скупове података који имају високе стопе недостајућих података, посебно у случајевима где постоји корелација између атрибута

Табела 6.1 Упоређивање алгоритама

Поред тога, битно је истаћи да постоји и низ нових техника за обраду података који недостају, као што су методе дубоког учења и Бајесове методе, које се показују као веома обећавајуће, али се још увек истражују. Ове методе могу бити посебно корисне за обраду великих и сложених скупова података.

У сваком случају, неопходно споменути и то да подаци који недостају могу садржати вредне информације које се не могу добити на други начин. Уз одговарајућу обраду, процена вредности које недостају може открити трендове, асоцијације и обрасце у подацима који би иначе остали неоткривени. Стога, необрађивање података који недостају може довести до губитка вредних информација и могућности за боље разумевање феномена који се проучава.

Такође, необрађени подаци који недостају могу довести до искривљених закључака и погрешних одлука, што може имати значајан утицај на пословне одлуке и правце политике. У економији, на пример, недостајући подаци о приходима или потрошњи могу довести до погрешних закључака о стању привреде или ефикасности политике. Слично томе, у медицинским истраживањима, недостајући подаци о здрављу пацијената могу довести до нетачних процена ефикасности лечења или дијагностичких процедура.

Стога би њихова обрада требало да буде обавезна фаза у анализи података, без обзира на врсту и количину података. Недостајући подаци могу бити присутни у било ком скупу података и важно је имати стратегију за руковање њима. Уз правилну обраду, подаци који недостају могу бити драгоцен извор информација и могу допринети бољем разумевању проучаваног феномена.

На крају, треба напоменути да руковање недостајућим подацима није крајњи циљ, већ само једна фаза у анализи података, сходно овоме је важно извршити све неопходне анализе и интерпретирати резултате на основу целокупног скупа података, укључујући и саме процене недостајућих вредности.

7 Литература

[1] *Best Practices in Data Cleaning_ A Complete Guide to Everything You Need to Do Before and After Collecting Your Data-SAGE Publications, Inc (2012)* Jason W. Osborne

[2] Третмани недостајућих података, Мирјана Облаковић, Валентина Соколовска, Бојана Динић

[3] Методе за обраду недостајућих података, доступно на: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUH%20Methods%20for%20Handling%20Missing%20Item%20Values%202018.pdf
(приступљено 9. марта 2023.)

[4] *Hot-deck*, доступно на: https://stats.stackexchange.com/questions/307339/hot-deck-imputation-it-preserves-the-distribution-of-the-item-values-how-c
(приступљено 10. марта 2023.)

[5] *Missing Data,* доступно на: https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf
(приступљено 10. марта 2023.)

[6] *Models that are resistant to missing values,* доступно на http://www.feat.engineering/models-that-are-resistant-to-missing-values.html
(приступљено 23. марта 2023.)