UNIVERZITET U NOVOM SADU

FAKULTET TEHNIČKIH NAUKA U NOVOM SADU

Dunja Vrbaški

Primena mašinskog učenja u problemu nedostajućih podataka pri razvoju prediktivnih modela

DOKTORSKA DISERTACIJA

NOVI SAD, 2020.

УНИВЕРЗИТЕТ У НОВОМ САДУ ● **ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА**
21000 НОВИ САД, Трг Доситеја Обрадовића 6

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, **РБР**:		
Идентификациони број, **ИБР**:		
Тип документације, **ТД**:		Монографска документација
Тип записа, **ТЗ**:		Текстуални штампани материјал
Врста рада, **ВР**:		Докторска дисертација
Аутор, **АУ**:		Дуња Врбашки
Ментор, **МН**:		др Купусинац Александар, др Дорословачки Ксенија
Наслов рада, **НР**:		Примена машинског учења у проблему недостајућих података при развоју предиктивних модела
Језик публикације, **ЈП**:		српски
Језик извода, **ЈИ**:		српски/енглески
Земља публиковања, **ЗП**:		Република Србија
Уже географско подручје, **УГП**:		АП Војводина, Нови Сад
Година, **ГО**:		2020.
Издавач, **ИЗ**:		Ауторски репринт
Место и адреса, **МА**:		Факултет техничких наука, 21000 Нови Сад, Трг Доситеја Обрадовића 6
Физички опис рада, **ФО**: (поглавља/страна/ цитата/табела/слика/графика/прилога)		7/112/70/14/25/3
Научна област, **НО**:		Електротехничко и рачунарско инжењерство
Научна дисциплина, **НД**:		Примењене рачунарске науке и информатика
Предметна одредница/Кључне речи, **ПО**:		машинско учење, недостајући подаци, предиктивни модели, вештачке неуралне мреже, случајне шуме
УДК		
Чува се, **ЧУ**:		Библиотека Факултета техничких наука у Новом Саду
Важна напомена, **ВН**:		
Извод, **ИЗ**:		Проблем недостајућих података је често присутан приликом развоја предиктивних модела. Уместо уклањања података који садрже вредности које недостају могу се применити методе за њихову импутацију. Дисертација предлаже методологију за приступ анализи успешности импутација приликом развоја предиктивних модела. На основу изнете методологије приказују се резултати примене алгоритама машинског учења, као метода импутације, приликом развоја одређених, конкретних предиктивних модела.
Датум прихватања теме, **ДП**:		
Датум одбране, **ДО**:		
Чланови комисије, **КО**:	Председник:	др Драган Иветић, редовни професор
	Члан:	др Јелица Протић, редовни професор
	Члан:	др Едита Стокић, редовни професор
	Члан:	др Горан Сладић, ванредни професор
	Члан, ментор:	др Александар Купусинац, ванредни професор
	Члан, ментор:	др Ксенија Дорословачки, ванредни професор

Потпис ментора

UNIVERSITY OF NOVI SAD ● **FACULTY OF TECHNICAL SCIENCES**
21000 NOVI SAD, Trg Dositeja Obradovića 6

KEY WORDS DOCUMENTATION

Accession number, **ANO**:		
Identification number, **INO**:		
Document type, **DT**:		Monographic publication
Type of record, **TR**:		Textual printed document
Contents code, **CC**:		Ph.D. thesis
Author, **AU**:		Dunja Vrbaški, M.Sc
Mentor, **MN**:		Kupusinac Aleksandar Ph.D, Doroslovački Ksenija Ph.D
Title, **TI**:		Application of machine learning to the problem of missing data in the development of predictive models
Language of text, **LT**:		Serbian
Language of abstract, **LA**:		Serbian/English
Country of publication, **CP**:		Republic of Serbia
Locality of publication, **LP**:		AP Vojvodina, Novi Sad
Publication year, **PY**:		2020
Publisher, **PB**:		Author's reprint
Publication place, **PP**:		Faculty of Technical Sciences, 21000 Novi Sad, Trg Dositeja Obradovića 6
Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes)		7/112/70/14/25/3
Scientific field, **SF**:		Electrical and Computer Engineering
Scientific discipline, **SD**:		Applied computer science and informatics
Subject/Key words, **S/KW**:		Machine learning, missing data, predictive modeling, artificial neural networks, random forests
UC		
Holding data, **HD**:		Library of the Faculty of Technical Sciences in Novi Sad
Note, **N**:		
Abstract, **AB**:		The problem of missing data is often present when developing predictive models. Instead of removing data containing missing values, methods for imputation can be applied. The dissertation proposes a methodology for analysis of imputation performance in the development of predictive models. Based on the proposed methodology, results of the application of machine learning algorithms, as an imputation method in the development of specific models, are presented.
Accepted by the Scientific Board on, **ASB**:		
Defended on, **DE**:		
Defended Board, **DB**:	President:	Dragan Ivetić, PhD, Full Professor
	Member:	Jelica Protić, PhD, Full Professor
	Member:	Edita Stokić, PhD, Full Professor
	Member:	Goran Sladić, PhD, Associate Professor
	Member, Mentor:	Aleksandar Kupusinac, PhD, Associate Professor
	Member, Mentor:	Ksenija Doroslovački, PhD, Associate Professor

Menthor's sign

Višnji i Urošu

Sadržaj

Lista slika

Lista tabela

Lista skraćenica

ML mašinsko učenje (eng. Machine Learning)

MD podaci koji nedostaju (eng. Missing Data)

LR linearna regresija (eng. Linear Regression)

DT stablo odluke (eng. Decision Tree)

RF slučajna šuma (eng. Random Forest)

ANN veštačka neuralna mreža (eng. Artificial Neural Network)

kCV k-slojna unakrsna validacija (eng. k-fold cross-calidation)

UVmd jednostruko nedostajanje (eng. Univariete missing data)

MVmd višestruko nedostajanje (eng. Multivariete missing data)

CC kompletni slučajevi (eng. Complete Cases)

MI višestruka imputacija (eng. Multiple Imputation)

1. Uvod

Prilikom rada sa podacima, bilo da se radi o analizi podataka i otkrivanju modela i odnosa među podacima ili se radi o konstrukciji prediktivnih modela, pojavljuje se često problem takozvanih *nedostajućih podataka* (eng. missing data).

Nedostajući podaci se odnose na sve informacije koje iz bilo kog razloga nisu evidentirane u polaznom skupu podataka, a bile bi od značaja za nastavak istraživanja. U struktuiranim podacima, za čiju reprezentaciju se koriste tabelarne reprezentacije, svaka informacija koja nedostaje u određenom polju tabele se smatra za nedostajuću. Nekompletne opservacije se mogu odbaciti i u daljem istraživanju se mogu koristiti samo one za koje su dostupni kompletni podaci. Međutim, još davno je pokazano da priroda nastajanja nedostajućih podataka može uticati na analizu i konačne rezultate [2] [3]. Pored grešaka u istraživanju, uklanjanjem podataka kod koji samo određeni deo nedostaje, smanjuje se i ukupna moć istraživanja jer imamo manje podataka na raspolaganju.

Vremenom su razvijane i korišćene metode koje omogućavaju takozvanu *imputaciju nedostajućih podataka* (eng. missing data imputation) i uz pomoć kojih se podaci koji nedostaju rekonstruišu ili procenjuju kako bi se omogućio dalji nastavak istraživanja. Razvoj metoda imputacije i njihovo istraživanje je predmet analize nedostajućih podataka i njime se mahom bave istraživači iz oblasti statističke analize.

U oblastima inženjerske primene statističkih metoda, kao što je mašinsko učenje i algoritamska izgradnja prediktivnih modela nad struktuiranim podacima, ovaj problem se često ne prepoznaje ili zanemaruje odbacivanjem podataka koji nedostaju ili implementacijom *ad hoc* izabranim algoritmima za imputaciju bez posebne analize samih nedostajućih podataka. Iako su vršena istraživanja u pravcu

korišćenja mašinskog učenja za imputaciju podataka retka su ona koja se detaljno bave uticajem nedostajućih podataka osim kao prevazilaženja prepreke u fazi pretprocesiranja.

Odatle nastaje i **predmet** istraživanja ove disertacije koji ima dvojaku prirodu: izučavanje razvoja prediktivnih modela mašinskog učenja nad struktuiranim podacima u prisustvu nedostajućih podataka kao i mogućnost korišćenja mašinskog učenja za imputaciju tih podataka.

Formirana su sledeća istraživačka pitanja:

Istrazivačko pitanje 1: Da li i kako podaci koji nedostaju utiču na razvoj prediktivnih modela mašinskog učenja?

Istrazivačko pitanje 2: Da li se same metode mašinskog učenja mogu iskoristi u rekonstrukciji podataka koji nedostaju?

Istraživačko pitanje 3: Da li se može formirati metodologija koja bi omogućila razvoj prediktivnih modela uz imputaciju nedostajućih podataka i validaciju korišćenja takve imputacije u odnosu na formiranje modela nad kompletnim skupom podataka?

Istraživačko pitanje 4: Da li se uz pomoć tako definisane metodologije može utvrditi da se neuralne mreže mogu iskoristi kao metod imputacije nedostajućih podataka za neke već poznate konkretne modele?

Na osnovu ovih pitanja formirane su sledeće hipoteze:

Hipoteza X_1: Veštačka neuralna mreža izvršava tačniju jednostruku imputaciju spoljašnjih, laboratorijskih vrednosti potrebnih pri izgradnji prediktivnog modela za procenu kardio-metaboličkog rizika.

Hipoteza X_2: Moguće je formirati smislenu i opštu metodologiju za analizu uticaja imputacije nedostajućih podataka pri izgradnji prediktivnih modela koja omogućava struktuiran, ponovljiv i sveobuhvatan dizajn eksperimenata neophodnih za ovu analizu.

Hipoteza X_3: Veštačke neuralne mreže se mogu iskoristiti za imputaciju nedostajućih podataka pri izgradnji prediktivnog modela za procenu metaboličkog sindroma zasnovanog na algoritmu slučajnih šuma.

U skladu sa navedenim istraživačkim pitanjima proizilaze i **ciljevi** istraživanja disertacije:

- istražiti i ukazati na značaj problema nedostajućih podataka, posebno u domenu razvoja prediktivnih modela mašinskog učenja,

- empirijski pokazati da se veštačka neuralna mreža može iskoristiti za imputaciju laboratorijskih vrednosti za potrebe izgradnje prediktivnog modela kardio-metaboličkog rizika,

- formirati precizan metodološki pristup koji omogućava analizu uticaja imputacije i formiranje modela bez posebne analize nedostajućih podataka i bez pretpostavki o njihovoj strukturi i mehanizmu nastajanja,

- eksperimentalno, kroz studiju slučaja, pokazati primenu i značaj date metodologije, testirajući neuralnu mrežu kao metod imputacije podataka koji se koriste za formiranje modela za procenu metaboličkog sindroma.

Značaj istraživanja i rezultata ove disertacije se ogleda u opštem i specifičnom doprinosu.

Opšti **doprinos** se odnosi na usmeravanje pažnje na jedan potproblem koji se pojavljuje prilikom razvoja modela mašinskog učenja i formalizaciju pristupa njegovom rešavanju usklađivanjem teorijskih koncepata iz izvornog mesta izučavanja ovog problema (statistika) i praktičnih obrazaca na mestu primene (mašinsko učenje) što je rezultiralo formiranjem dve razvojne metodologije za pristup problemu nedostajućih podataka pri razvoju prediktivnih modela mašinskog učenja.

Specifični **doprinos** predstavljaju rezultati eksperimenatalnih istraživanja. U toku rada na ovom doktoratu su objavljene dve studije u međunarodnim časopisima [1, 4] dok su u disertaciji prikazani rezultati studije koja, kroz primenu ranije navedene formalizacije, utvrđuju da se neualne mreže mogu koristiti kao metod imputacije nedostajućih podataka prilikom realizacije prediktivnog modela preko slučajnih šuma za procenu metaboličkog sindroma.

Disertacija ima sledeću **strukturu**.

Na početku, u poglavlju 2, je dat kratak osvrt na kontekst iz kog je proisteklo ovo istraživanje. Ukazano je na razlike između dve naučne zajednice i mesto koje problem nedostajućih podataka zauzima u njima.

U poglavlju 3 je prikazana terminologija i često korišćeni pojmovi koji se odnose na problem nedostajućih podataka. Takođe, predstavljen je i pojam imputacije, kao opšti pojam koji predstavlja procese rekonstrukcije naspram procesa

eliminacije podataka koji nisu kompletni uz odgovarajući pregled dela literature koja se odnosi na upoređivanje različitih metoda imputacije.

U poglavlju 4 je predstavljena oblast mašinskog učenja koja je našla svoje mesto u primeni u problemu nedostajućih podataka, pre svega kod imputacija. Nakon osvrta na metode mašinskog učenja koje su korišćene u eksperimentalnom delu istraživanja dat je i pregled literature koja se odnosi na primenu mašinskog učenja u problemu nedostajućih podataka. Na kraju su, sumarno, prikazani rezultati istraživanja, prihvaćeni za objavljivanje [1], koji se odnose na potvrđivanje hipoteze X_1 odnosno na utvrđivanje ponašanja veštačkih neuralnih mreža kao metoda imputacije podataka potrebnih za izgradnju modela predikcije kardio-metaboličkog sindorma.

Poglavlje 5 prikazuje, prvo, značaj razmatranja problema neodostajućh podataka u obrnutom kontekstu. Razmatra se kako podaci koji nedostaju utiču na razvoj samih prediktivnih modela mašinskog učenja: kad se ovaj problem javlja, ko je odgovoran za njegovo rešavanje, kako mu pristupiti i šta očekivati u literaturi? Zatim se daje pregled opštih i tehničkih pojmova i procesa neophodnih za definisanje metodologije za analizu uticaja imputacija prilikom razvijanja prediktivnih modela, nakon čega sledi i sam prikaz dva algoritma što se odnosi na potvrđivanje hipoteze X_2. Uz svaki algoritam su date posebne i opšte napomene vezane za realizaciju i koje obezbeđuju inženjerima koji se bave primenom mašinskog učenja, ispravnu analizu uticaja bez potrebe za značajnom ekspertizom iz statistike i zasebnom analizom nedostajućih podataka.

U poglavlju 7 je jedan od algoritama primenjen na proces razvoja konkretnog prediktivnog modela objavljenog u [4]. Kroz niz različitih slučajeva nedostajućih podataka analiziran je uticaj imputacija realizovanih preko veštačkih neuralnih mreža na formiranje finalnog prediktivnog modela. Utvrđeno je da se za dati model može, ukoliko je potrebno, izvršiti ovakva vrsta imputacija za različite strukture i mehanizme nastajanja nedostajućih podataka. Metod imputacije pomoću neuralnih mreža se pokazao kao neosetljiv na mehanizme i strukture nedostajanja u smislu poboljšanja performansi finalnog prediktivnog modela u odnosu na korišćenje podskupa kompletnih podataka čime je potvrđena hipoteza X_3.

Na kraju su, u poglavlju 8, diskutovani izvedeni zaključci i predložene su smernice za nastavak istraživanja i budući rad.

2. Problem nedostajućih podataka i dve naučne zajednice

Pre nego što budu predstavljeni problemi koji se javljaju prilikom razvoja prediktivnih modela mašinskog učenja, osvrnuću se na razliku između statistike i mašinskog učenja jer smatram da se tu nalazi odgovor na pitanje zašto, u okviru primene mašinskog učenja, problemu nedostajućih podataka nije posvećena prevelika pažnja, delom u literaturi, a najviše u domenu praktične primene?

Treba napomenuti čestu dilemu o razlici i preseku statistike i mašinskog učenja koja neretko započinje pitanjem da li jednostavna linearna regresija zaista predstavlja algoritam mašinskog učenja. Ovde se nećemo baviti tom raspravom već samo skrećem pažnju na nju jer smatram da je od značaja za predmet ove disertacije. Naime, problem podataka koji nedostaju se upravo nalazi u samom preseku statistike i mašinskog učenja. Njegovo glavno fundamentalno, teorijsko uporište je u statistici. Sa druge strane, razvoj prediktivnog modela uz prisustvo nedostajućih podataka se nalazi potpuno u domenu primene mašinskog učenja pa otud i i nastaje različit odnos prema ovom problemu.

Flek, jedan od začetnika filozofije i sociologije nauke, sa kojim se u nekim delovima i Kun slagao [5], u *Nastanku i razvoju naučne činjenice* [6], iznosi svoju tezu da je znanje kolektivno, u smislu da svaka istina, teorija i misao pripada jednoj odgovarajućoj misaonoj kulturi koja, opet, odgovara jednoj zajednici. Ono što je tačno i značajno u jednoj (naučnoj) zajednici ne mora biti tačno i važno u drugoj [7].

Još davno je Breiman [8] ukazao na suštinsku razliku između dva domena, statistike i algoritamskog modelovanja. U statistici smo, obično i pre svega, za-

interesovani za analizu podataka, za to da opišemo prirodu uzorka i populacije iz koje podaci nastaju. U tom smislu, pretpostavljamo da postoji i želimo da otkrijemo model koji opisuje podatke. S druge strane, u primeni metoda mašinskog učenja to nije primarni cilj. Najčešće nismo zainteresovani za takvo objašnjavanje ulaznih podataka i njihovih raspodela već prvenstveno za njihovo korišćenje u razvoju i validaciji modela i rezultate koje naš model ostvaruje nad njima [9, 10]. U statistici smo usredsređeni najviše na ispravnost zaključivanja (hipoteze, testovi, analiza,...) dok smo kod primene mašinskog učenja znatno više usredsređeni na ispravnost procedure (preprilagođavanje, usložnjavanje, transformacije polaznog skupa,...).

Iz tog razloga se može pretpostaviti da manja zainteresovanost za značaj problema nedostajućih podataka kod inženjera i naučnika koji se bave primenom mašinskog učenja nije namerna već prostekla isključivo iz pravca i ciljeva istraživanja i njihove ekspertize. S druge strane, kao što je rečeno u [11], iako ne moramo nužno znati da implementiramo algoritme mašinskog učenja potrebno je upoznati sve procese koji predstavljaju njihovu osnovu kako bismo mogli da ih uspešno i valjano primenjujemo, a ti procesi se često oslanjaju na statistiku i druge matematičke oblasti.

Iako se danas, u odnosu na 2001. godinu, dve zajednice znatno više razumeju i koriste prednosti i dostignuća obe nauke, problem nedostajućih podataka je, na neki način, i dalje ostao u onoj ravni neprepoznavanja, ako ništa drugo, onda njegovog značaja za dalji tok razvoja finalnih modela.

Tako je nastala i motivacija za realizacijom istraživanja predstavljenog u disertaciji, koje bi, svojim delom, predstavljalo sredstvo da se problem prepoznat u jednoj misaonoj kulturi (statistika) približi drugoj (primena mašinskog učenja) koristeći jezik, porocese i ciljeve koji su u njoj poznati i prihvaćeni.

3. Podaci koji nedostaju

3.1 Problem nedostajućih podataka

Podaci koji nedostaju (u daljem tekstu MD) predstavljaju čest problem prilikom statističkih obrada podataka, ali i prilikom primene algoritama mašinskog učenja nad njima. U mnogim bazama podataka određene informacije nedostaju, u većoj ili manjoj meri. Razlozi za nedostajanje mogu biti različiti: podaci mogu biti nedostupni, pogrešno evidentirani, uništeni ili izgubljeni.

Prema [12] definicija nedostajućih podataka bi mogla biti sledeća.

Definicija 1 ***Nedostajući podaci*** *su neevidentirane vrednosti koje bi bile od značaja za analizu da su evidentirane; drugim rečima, nedostajuća vrednost skriva smislenu vrednost.*

Ukoliko se upravljanju ovakvim podacima, najčešće u toku pretprocesiranja, ne pristupi na odgovarajući način može doći do određenih problema u nastavku istraživanja i formiranju neispravnih finalnih modela [2, 3, 13, 14, 15, 16].

Prilikom istraživanja, u ranoj fazi, ukoliko postoje podaci koji nedostaju možemo uočiti nekoliko pitanja odnosno potproblema:

- Šta raditi sa podacima koji nedostaju?
- Da li i kako izbor tih akcija utiče na performanse realizovanih modela i rezultate istraživanja?

Za tetman podataka kojima nedostaje određeni deo, zavisno od modela, problema, ali i onog koji vrši zadatak, može se izabrati:

- isključivanje MD u celini - svi zapisi koji imaju neki podatak koji nedostaje se isključuju (brišu) iz polaznog skupa podataka,
- isključivanje promenljivih koje sadrže MD,
- kodiranje informacije o nedostajanju odnosno uključivanje nove promenljive koje predstavlja indikator nedostajanja,
- popunjavanje novim vrednostima koje su generisane određenim metodama,
- kombinacija ovih metoda.

Izbor akcija može uticati na dalji tok istraživanja o čemu će biti reči, u više navrata, u nastavku ove disertacije.

3.2 Mehanizmi nedostajanja

Jasno je da podatke koji nedostaju ne možemo nikako poznavati, a time ni analizirati. Međutim, možemo izučavati prirodu njihovog nastanka i u kakvoj su eventualnoj vezi sa sa poznatim, dostupnim podacima. Na osnovu toga obično razlikujemo tri tipa MD [12, 17, 18]:

- **Missing at Random (MAR)** - kada je nedostajanje uslovljeno drugim dostupnim podacima,
- **Missing not at Random (MNAR)** - kada je nedostajanje uslovljeno samim podacima koji nedostaju,
- **Missing Completely at Random (MCAR)** - kada nedostajanje podataka nije ni u kakvoj relaciji sa ostalim podacima niti sa samim sobom.

Ovu terminologiju je uveo Rubin još 1976 [17] koji se i smatra začetnikom analize nedostajućih podataka. Iako mnogi autori zameraju pomalo nezgrapnom izboru termina, ovi nazivi su se zadržali do danas i korišćeni su u ovoj disertaciji.

Ako za podatke Y sa M označimo nedostajanje, a sa θ nepoznate parametre onda se distribucija podataka koji nedostaju po ovim mehanizmima može zapisati na sledeći način:

$$P(M|Y, \theta) = P(M|Y_{obs}, \theta) \tag{MAR}$$

$$P(M|Y, \theta) = P(M|Y_{miss}, \theta) \tag{MNAR}$$

$$P(M|Y, \theta) = P(M|\theta) \qquad \text{(MCAR)}$$

gde *obs* označava dostupne podatke (eng. observed), a *miss* podatke koji nedostaju (eng. missing).

Pokazano je, u statističkom modelovanju, da jedino MCAR podaci daju nepristrasne procene parametara i to ukoliko je količina ovih podataka mala (< 5%) [19]. Iako MCAR mehanizam možemo potvrditi sa određenom tačnošću koristeći Litleov MCAR test [20], za MAR i MNAR podatke ne možemo formalno utvrditi o kom se mehanizmu radi. Takođe, uvođenje pretpostavke o mehanizmima bez posebne analize ili znanja i potvrda o njima može takođe dovesti do različitih rezultata [21, 22].

3.3 Struktura nedostajanja

Podaci mogu nedostajati u različitim oblicima. Nepoznate vrednosti se mogu pojavljivati samo kod jednog ulaznog parametra ili kod više. U tom slučaju razlikujemo jednostruko i višestruko nedostajanje (eng. univariete and multivariete missing, u daljem tekstu UVmd i MVmd). Kod MVmd se mogu pojavljivati različita, dodatna pravila ili strukture nedostajanja. Na primer, mogu se pojavljivati grupe promenljivih kod kojih vrednosti uvek nedostaju zajedno.

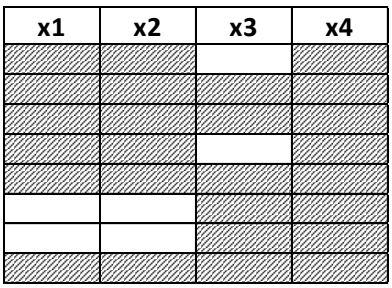
Ovi oblici nedostajanja u podacima se često opisuju pomoću koncepta koji se zove *patern nedostajanja*. Jedan patern nedostajanja predstavlja informaciju o jednom mogućem rasporedu prisutnih i nedostajućih podataka za neka posmatranja. Broj paterna je konačan i može biti najviše jednak broju podskupova promenljivih bez praznog i celog skupa.

Prilikom implementacija, za reprezentaciju paterna je pogodna predstava preko vektora dužine n čiji su elementi 0 i 1 gde 0 označava da određena vrednost za datu promenljivu nedostaje, a 1 da je vrednost poznata.

Definicija 2 *Neka je* $X = \{x_1, x_2, ..., x_K\}$ *skup promenljivih polaznog skupa.* ***Skup paterna nedostajanja*** *je skup* $P = \{p_1, p_2, ..., p_k\}$ *gde je* K *broj promenljivih,* k *je broj paterna,* $p_i = (m_1, m_2, ..., m_K)$*, a*

$$m_i = \begin{cases} 1, & \textit{vrednost postoji za } x_i \textit{ u podskupu koji odgovara paternu } p_i \\ 0, & \textit{vrednost nedostaje} \end{cases} \tag{3.1}$$

Na slici 3.1 je prikazana šema jednog skupa podataka sa 4 promenljive gde označena polja predstavljaju sve poznate vrednosti dok prazna polja označavaju vrednosti koje nedostaju.



Slika 3.1: Primer dva paterna nedostajanja u jednom skupu

U ovom primeru, gde je $K = 4$, se pojavljuju dva paterna nedostajanja ($k = 2$):

- $p_1 = (1, 1, 0, 1)$ i
- $p_2 = (0, 0, 1, 1)$.

Napominjemo da skup X predstavlja ceo polazni skup promenljivih jer podaci mogu nedostajati i za prediktore i za izlazne promenljive i za sve dodatne promenljive koje potencijalno ne moraju učestvovati u nekom budućem modelu.

U literaturi se nekad, kao što je kod [12], pod paternom nedostajanja smatra čitava matrica dimenzija istih kao polazni skup gde su sa 0 i 1 označene vrednosti koje su prisutne odnosno koje nedostaju za svaku vrednost polaznog skupa. Ta matrica se onda naziva i *matrica indikatora nedostajanja*. U uvoj disertaciji se zadržava prethodno navedena definicija 2 radi jasnijeg, kasnijeg, povezivanja sa implementacijom.

Paterni, ovako definisani, predstavljaju čisto formalni opis strukture MD i nemaju vezu sa mehanizmima nedostajanja. Međutim, obično se prilikom analize, paterni posmatraju u sklopu nekog mehanizma nedostajanja pa u tom kontekstu

uvodimo ovde još jedan koncept, a to je *scenario nedostajanja*. Pod jednim scenarijom ćemo smatrati polazni skup sa čitavim skupom paterna zajedno sa odgovarajućim mehnizamima nedostajanja. Ovaj pojam uvodimo, pre svega, zbog jednostavnije klasifikacije i notacije analiziranih konkretnih slučajeva nedostajućih podataka u nastavku istraživanja.

Definicija 3 ***Scenario nedostajanja*** *predstavlja jedan konkretan slučaj nedostajanja podataka u nekom skupu određen skupom paterna i mehanizmima nedostajanja pod kojima se pojavljuju nedostajući podaci.*

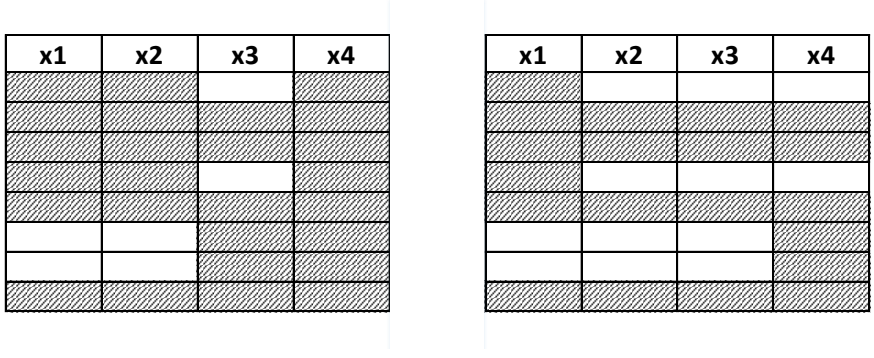
Praktično, kad odlučimo ili utvrdimo postojanje neke zavisnosti među MD možemo formirati odgovarajući scenario. Na primer, možemo reći: posmatramo scenario u kom tri promenljive nedostaju zajedno, po MAR mehanizmu i u znatnoj meri.

Ovo nas dovodi i do koncepta *količine* MD odnosno mere koliko podataka nedostaje.

Definicija 4 ***Količina nedostajanja*** *predstavlja procenat opservacija u skupu podataka koje sadrže nedostajuću vrednost za bar jednu promenljivu.*

Potrebno je skrenuti pažnju na terminologiju jer ovakva definicija ukazuje na to da dva skupa sa istom količinom MD mogu imati značajno različitu ukupnu količinu podataka koji nedostaju. Na primer, za zadatu količinu MD, skup u jednom scenariju može imati samo jednu promenljivu za koju nedostaju vrednosti dok se u drugom scenariju mogu pojaviti nedostajući podaci za sve promenljive.

Na slici 3.2 su prikazana dva istovetna skupa, sa istim promenljivama, u kojima se MD pojavljuju u istoj količini od 50% iako u drugom skupu postoji više ukupno nepoznatih vrednosti.



Slika 3.2: Primer iste **količine** nedostajanja (definicija 4) u skupovima sa istom vrstom podataka.

3.4 Imputacija podataka

Kao što je ranije navedeno, umesto odbacivanja podataka koji sadrže MD, može se pristupiti procesima imputacije podataka kada se nekom izbaranom metodom kompletiraju podaci koji nedostaju. Ustaljeni naziv za ovaj proces je *imputacija* (eng. imputation).

U literaturi se obično pravi razlika između jednostavnijih i neprednijih tehnika imputacije. Jednostavnije tehnike podrazumevaju imputaciju na osnovu nekog parametra kao što je srednja vrednost, korišćenje indikatora koji ukazuju na nedostajanje, primenu nekih jednostavnijih prediktivnih metoda kao što je linearna regresija ili takozvane hot-deck imputacije.

U naprednije metode se ubrajaju metode koje su kasnije razvijene, ali koje su: često računski zahtevnije, obično pretpostavljaju MAR mehanizam i, što je najveće ograničenje, skoro uvek zahtevaju dobro poznavanje i razumevanje statističkih metoda kao i teoriju MD. Ukoliko se radi o MNAR podacima, kada dostupni podaci ne sadrže nikakvu informaciju o nedostajanju, takvi postupci često ne garantuju ispravnost rešenja. Zato je uglavnom neophodno eventualne MNAR mehanizme posebno modelovati ili izvršti analizu osetljivosti na različite mehanizme nedostajanja. U literaturi su najpoznatije metode višestruke imputacije (eng. Multiple Imputation, u daljem tekstu MI) i maksimalne verodostojnosti (eng. Maximum Likelihood). Obe metode su razvijane i najviše istraživane u domenu statistike u okviru postupaka za ocene statističkih parametara.

Metoda maksimalne verodostojnosti zapravo ne vrši imputacije već na osnovu MD i dostupnih podataka vrši procenu parametara statističkog modela i odnosi se na podatke koji imaju normalnu raspodelu. Stoga neće biti dalje razmatrana jer izlazi van opsega disertacije zbog drugačijih ciljeva tretmana MD pri razvoju predektivnih modela mašinskog učenja, o čemu će biti reči u nastavku. Višestruka imputacija (MI) takođe za cilj ima procenu parametara modela koji opisuje podatke, ali na ovaj metod se treba osvrnuti jer u svom postupku ipak vrši određene imputacije.

Metod višestruke imputacije je, bar među istraživačima koji su detaljnije upoznati sa problemom MD, zasigurno najpopularniji metod za imputaciju. Predložio ga je i formalizovao Rubin [17]. Ideja višestruke imputacije je veoma jednostavna. U osnovi ima za cilj da na neki način modeluje nesigurnost, odnosno neznanje, o podacima koji nedostaju što druge jednostavne metode nisu ostvari-

vale. Prvi korak postupka podrazumeva da se svaka vrednost koja nedostaje, na osnovu određenih statističkih modela, popunjava sa proizvoljno, konačno (m) različitih vrednosti čime se dobija m novih, popunjenih skupova. Razlika između tih skupova bi, praktično, trebala da predstavljaja navedenu nesigurnost koju donose MD. Dalja statisti;ka analiza podataka se vrši na svakom od tih m novih, kreiranih skupova, a konačni rezultati se sumiraju u jedan, na osnovu pravila koje su precizirane takozvanim Rubinovim pravilima. Definišući pravila, on je pokazao i da se ovakvom procedurom dobijaju validniji statistički rezultati, odnosno nepristrasne porcene parametara, ukoliko se radi o MAR mehanizmu i u odnosu na druge metode.

Ovde se odmah ukazuje problem koji se javlja u domenu primene mašinskog učenja i koji se tiče korišćenja navedene metode, a to je upravo kontekst i cilj u kom se ova procedura razvijala, primenjivala i istraživala, a to je statistika. Većina literature se upravo odnosi na istraživanje višestruke imputacije primenjene u okvirima analize modela zasnovanih na linearnoj ili logističkoj regresiji i analizom koeficijenata dobijenih modela radi utvrđivanja funkcije koja modelira podatke. Za razliku od ovakve analize, kod primene mašinskog učenja nad struktuiranim podacima nam je potrebna jedna nova, konačna, imputirana vrednost koja dalje može, kao deo skupa, nastaviti u proceduru razvoja prediktivnog modela. Ako se dodatno uzme u obzir da je sama metoda višestruke imputacije računski složena, da zahteva određeno predznanje i pretpostavlja mehanizam nedostajanja onda postaje opravdano da se umesto nje razmotri neka druga metoda.

Sve metode imputacije su znatno i dugo istraživane i postoji ogroman korpus literature koji se odnosi na analizu i unapređenje metoda, prikaza primena u različitim oblastima, upoređivanja metoda u različitim slučajevima i utvrđivanju njihove ispravnosti. DS tim da se pod ispravnim smatra ono što je ispravno u domenu inferencijalne statistike, što znači da se na kraju dobija model koji dobro predstavlja nesigurnost koju donose MD, a opet omogućava dobijanje dobrih parametara modela koji opisuju polazne podatke. Cilj razvoja prediktvnog modela mašinskog učenja je malo drugačiji jer se vrši potraga za modelom koji dobro vrši predikciju uz prisustvo MD bez značaja dodatne analize modela ulaznih podataka.

3.5 Pregled stanja u oblasti

U ovom poglavlju je dat prikaz dela istraživanja koja su bila usmerena na upoređivanje različitih metoda imputacije i predstavlja podskup literature koja je kon-

sultovana u izradi ove disertacije. Navedeni su specifični primeri studija kako bi se stekla opšta slika o pravcima izučavanja u ovoj oblasti pa ovaj pregled svakako ne predstavlja sveobuhvatnu literaturu koja se odnosi na ovu temu.

Pre prikaza studija treba napomenuti da se u značajnom broju istraživanja upoređuju različite metode imputacije sa slučajem kada se u analizi koriste samo dostupni podaci, odnosno oni kod kojih nema MD. Ovo se obično naziva *analiza kompletnih slučajeva* (eng. complete-case, u daljem tekstu CC).

Donders et al. u [23] daju lep uvod o tretmanu MD i na primeru koji podrazumeva jedan MAR scenario i sintetičke podatke ukazuju na probleme koje nastaju prilikom jednostrukih i višestrukih imputacija sve u cilju objašnjavanja promena u koeficijentu i standradnoj grešci logističke regresije.

Marshall et al. u [24] upoređuju generisane podatke koji imitiraju raspodelu koja se sreće kod predikcije raka dojke. Ispitivanjem sva tri mehanizma i nekoliko količina nedostajanja su upoređivali CC slučaj, jednostruku i višestruku imputaciju (MI) korišćenjem statističkih metoda. Pokazali su da prilikom analize koeficijenata regresije MI treba da bude izabrani metod, ali pod uslovom da postoji manja količina MD i da mehanizam nedostajanja nije MNAR. Slično, Ali et al. u [25], ponovo utvrđuju da se MI može preferirati u odnosu na CC, ali da rezultati nisu značajno različiti. Ono što su oni naglasali, što nije uvek slučaj u literaturi, a to je značaj razmatranja i uključivanja izlazne promenljive u model imputacije.

Stavseth et al. u [26] upoređuju šest MI metoda imputacije za kategorijalne vrednosti u uptnicima. Podaci su iz CC podskupa uklanjani u različitom obimu. Pokazali su da MI daje bolje procene nego CC, ali da se poboljšanja smanjuju sa porastom količine MD. Međutim, testiran je samo jedan MAR slučaj, a analiza rezultata je opet vršena u odnosu na promene koeficijenata logističke regresije.

Masconi et al. u [27] su prilikom upoređivanja metoda imputacije pri razvoju prediktivnog modela za dijabetes utvrdili da je CC pristup davao lošije rezultate, ali da nije postojala značajnija razlika između jednostruke i višestruke imputacije.

Choi et al. u [28] se analizirali efekte imputacije u *propensity score* analizi [29] koja uključuje i postojanje ometajućih (eng. confounding) promenljivih. Koristeći generisane podatke su testirali uticaj imputacije u prisustvu i bez postojanja tog faktora i došli su do zaključka da čak i CC slučajevi mogu dati dobre rezultate, da MI, ako se koristi sa predefinisanim metodama imputacije, nije od velike pomoći a da MNAR mehanizam iznova stvara probleme.

Hughes et al. u [30] opet potvrđuju da MI ne predstavlja univerzalni odgovor na svaki MD problem razmatrajući dva modela, regresioni i logistički, i uzimajući u obzir i dodatne promenljive u analizi različitih mehanizama.

U svim navedenim istraživanjima zaključuje se, između ostalog, da još uvek ne postoji univerzalni alat koji omogućava uvek ispravnu imputaciju i da je neophodno pažljivo pristupiti analizi ovog problema. Utvrđeno je, takođe, i da MI metodologija nije svemoguća, ali da je svakako treba razmotriti ukoliko je cilj istraživanja modelovanje podataka.

4. Primena mašinskog učenja u problemu nedostajućih podataka

Mašinsko učenje je, kao i u mnogim drugim oblastima, našlo svoju primenu i u problemu nedostajućih podataka. Stoga su u ovom poglavlju izneti osnovni pojmovi i korišćene metode mašinskog učenja kao i okviri istraživanja i ostvareni rezultati u domenu njegove primene u problemu nedostajućih podataka.

4.1 O mašinskom učenju

Mašinsko učenje (eng. machine learning, u daljem tekstu ML), kao podoblast veštačke inteligencije, predstavlja disciplinu izučavanja algoritama i računarskih sistema koji na osnovu ulaznih podataka i bez dodatnih instrukcija izvršavaju određene zadatke kao sto su predikcija ili klasifikacija. Jednostavnije, metode mašinskog učenja umeju da spoznaju i nauče određeno znanje iz ulaznih podataka.

Danas je prva asocijacija na veštačku inteligenciju i mašinsko učenje oblast primene koja se odnosi na velike količine podataka koji su često nestruktuirani i gde je učenje nenadgledano ili potkrepljujuće. U ovoj disertaciji je razmatrano nadgledano učenje nad struktuiranim, tabelarnim podacima čije su vrednosti prosti tipovi podataka.

Kad govorimo o nadgledanom mašinskom učenju pretpostavljamo da imamo neki ulaz, izlaz i funkciju koja opisuje vezu između njih. Obično se zapisuje u vektorskom obliku kao $f(X) = Y$ gde X predstavlja vektor ulaznih vrednosti, Y izlazne vrednosti, a f predstavlja funkciju zavisnosti. Cilj mašinskog učenja je

da, koristeći ulazne i izlazne podatke, pronađe takvu funkciju f' koja će uspešno aproksimirati polaznu funkciju f koja nam je nepoznata. Pod uspešnim smatramo da će rezultujuća funkcija koju je algoritam mašinskog učenja pronašao, osim da ima što manju grešku na obučavajućem skupu, dobro generalizuje početni problem odnosno da isto tako dobro opisuje i nove, nepoznate podatke.

4.2 Prediktivne metode mašinskog učenja

U nadgledanom mašinskom učenju obično razlikujemo dve vrste modela u odnosu na vrstu izlaznih vrednosti. Zavisno od toga da li se radi o kvantitativnoj ili kvalitativnoj vrednosti modele delimo na: regresione i klasifikacione prediktivne modele [31]. Do danas je razvijen veliki broj algoritama, obe vrste, s tim da se neki od njih mogu koristiti za obe vrste zadataka uz manje izmene. U okviru ovog istraživanja su korišćene sledeće metode mašinskog nadgledanog učenja: linearna regresija, stabla odluke, slučajne šume i veštačke neuralne mreže.

Linearna regresija (eng. *Linear Regression*, u daljem tekstu LR) predstavlja jednostavan statistički model koji pokušava da definiše linearnu vezu između nezavisnih i zavisnih promenljivih. Iako LR nije u mogućnosti da dobro modelira komplikovanije nelinearne veze, za razliku od ostalih metoda, ovaj model je u potpunosti objašnjiv i često se uzima kao reper u istraživanjima koja se bave komparacijom metoda. Takođe, popularan je i u medicinskoj literaturi i poznat medicinskim stručnjacima. Zato smatramo da je poželjen u okvirima ovog istraživanja s obzirom da su eksperimenti vršeni nad podacima iz ovog domena.

Stablo odluke (eng. *Decision Tree*, u daljem tekstu DT) je metod koji za ulaznu vrednost daje izlaznu vrednost na osnovu rezultata dobijenih nizom određenih odluka. Sastoji se iz čvorova u kojima se donose odluke, grana koje predstavljaju rezultate odluka i listova koji predstavljaju konačne vrednosti. Postoje klasifikaciona i regresiona stabla odluke zavisno od tipa izlaznih vrednosti (listova). U odnosu na linearne modele, umeju uspešnije da modeluju nelinearne zavisnosti, a uglavnom i dalje zadržavaju visok stepen interpretabilnosti i objašnjivosti.

Slučajne šume (eng. *Random Forest*, u daljem tekstu RF) predstavljaju model kog čini skup stabala odluke. Najčešće su ta stabla kreirana upakivanjem odnosno slučajnim izborom trening podataka (eng. bagging) i slučajnim izborom podskupa promenljivih. Prilikom predikcije, svako stablo vrši predikciju i izglasavanjem biva izabrano konačno rešenje ili se, u slučaju problema regresije, za rezultat uzima, re-

cimo, prosečna vrednost. RF su uvedeni kako bi se smanjilo preprilagođavanje (eng. overfitting) jednog DT modela. S druge strane dobijen je manje objašnjiv model što u nekim oblastima može predstavljati problem prilikom prihvatanja algoritamskog rešenja.

Veštačka neuralna mreža (eng. *Artificial Neural Networks*, u daljem tekstu ANN) predstavlja metod čija je ideja zasnovana na pokušaju imitacije bioloških neuralnih mreža. Sastoji se iz skupa neurona i veza između njih preko kojih se na određeni način prenose "signali". Postoje ulazni i izlazni neuroni. Ulazni neuroni predstavljaju ulazne podatke, a izlazni neuroni (kod nadgledanog učenja) mogu predstavljati očekivani izlaz odnosno rezultat modela. Prilikom obučavanja se meri greška u odnosu na očekivani rezultat i vrše se izmene u mreži sve dok se ne postigne određeni nivo tačnosti. Neuroni mogu biti podeljeni i na slojeve (eng. deep learning [32]) kako bi se bolje modelirala nelinarna funkcija između ulaznih i izlaznih neurona. ANN modeli su veoma moćni i koriste se podjednako i za nadgledano i nenadgledano učenje. Mnogi nelinearni problemi nad nestruktuiranim podacima kao što su slike i tekstovi u slobodnoj formi i pogotovo kad ih ima u velikoj količini (Big Data) se danas uspešno rešavaju primenom ANN.

4.3 Pregled stanja u oblasti

Što se tiče korišćenja metoda mašinskog učenja u domenu MD, u nastavku je dat osvrt na nekoliko istraživanja koja su od značaja za istraživanja u okviru ove disertacije.

Pesonen et al. u [33] su razmatrali kako imputacija MD utiče na tačnost prediktvnog modela upoređujući nekoliko metoda. Za istraživanje su korišćeni realni podaci, bez posebnog osvrtanja na mehanizme. Upoređivanje je vršeno za imputaciju MD jedne promenljive koja je imala značajniju količinu MD. Utvrđeno je da se ANN bolje, ali slično ponaša kao druge metode. Međutim, druge dve stvari su značajne u ovom radu. Prvo, autori su naglasili da su u prethodnim radovima razvijali finalni model samo nad kompletnim podacima, a ovakvi navodi nisu čest slučaj u literaturi. Drugo, iskoristili su analizu MD, osim za merenje performansi finalnog modela, i za utvrđivanje značaja same promenljive na finalni model.

Silva-Ramirez et al. u [34] su uradili obimno istraživanje imputacije preko ANN nad 15 realnih i sintetičkih skupova različite veličine. Takođe, testirali su različite algoritme učenja za ANN i uporedili su ih sa nekim jednostavnim metodama

imputacije. Utvrdili da ANN ostvaruje generalno najbolje rezultate. Ono što je najznačajnije u ovom radu, pored precizno prikazane metodologije, je i zaključak da ANN znatno bolje rešava MD problem kada postoje kategorijalne promenljive dok u skupovima gde su sve vrednosti kvantitativne, performanse ANN i jednostavnijih modela su sličnije što ide u prilog ovim drugim zbog složenosti implementacije i izvršavanja.

Beaulieu et al. u [35] su upoređivali MI metod (sa različitim modelima imputacije) sa neuralnom mrežom sa 1, 2 i 3 skrivena sloja i različitim brojem neurona koja je implementirana tako da uzima u obzir i indikator nedostajanja. Testirane su različite količine nedostajanja i MCAR i MNAR mehanizmi. Nad imputiranim skupovima je kreiran mode slučajne šume pa su se na osnovu performansi modela utvrđivale i performanse imputacija. ANN su pokazale najbolje performanse i utvrdili su važan zaključak a to je da tačnost imputacije ne mora biti u korelaciji sa performansama finalnog modela, ali da ipak najtačniji model imputacije omogućuva i najtačniji finalni model.

Leke et al. u [36, 37] su istraživali primenu neuralnih mreža zajedno sa genetskim algoritmima za optimizaciju testirajući svoja rešenja i predstavljajući teorijski okvir i metodologiju za pristup rešavanju MD. Prilikom simulacija testirani su MCAR i MAR mehanizmi nad skupom sa 10% nedostajanja, a rezultati su upoređivani sa jednostavnom neuralnom mrežom kao metodom imputacije. Iako su postignuti rezultati dobri ostao je problem znatnog povećanja složenosti što dalje uslovljava povećanje vremena izvršavanja i manje interpretabilnosti. Rezultati koji prethode ovim istraživanjima sa sličnim ciljevima se mogu videti i u [38, 39, 40].

Za razliku od prethodnih istraživanja, koja su nama bila značajna, ima studija koje se bave primenom ML metoda za imputaciju, ali i za drugačije vrste podataka u odnosu na naše. Tako su Duan et al. u [41] pokazali da se deep learning model može uspešno iskorisiti za imputaciju loših ili nedostajućih podataka koji se prikupljaju sa senzora za analizu saobraćaja dok su Kim et al. u [42] istraživali primenu ANN i DT kod imputacije time-series meteoroloških podataka.

Sva navedena istraživanja sugerišu da se ANN i drugi modeli ML mogu, i treba, razmotriti kao potencijalne metode imputacije. Međutim, mahom su istraživanja usmerena na istraživanje prediktora, a kao što će biti izloženo u nastavku, postoji određena podgrupa problema gde je neophodno vršiti imputaciju vrednosti spoljašnjih promenljivih na osnovu kojih se računaju izlazne vrednosti.

Takođe, često su analizirani problemi sa fiksnim mehanizmima nedostajanja, količinama, paternima ili su razmatrani sa konačnim ciljem gde se izgrađuju statistički deskriptivni modeli.

Iako imputacija predstavlja najčešći cilj upotrebe mašinskog učenja u problemu nedostajućih podataka postoje i druge primene. Na primer, Tierney et al. u [43] koriste slučajne šume da prepoznaju strukturu nedostajanja i promenljive koje utiču na nedostajanje bez otkrivanja o kom se mehanizmu radi. Razzaghi et al. u [44] kreiraju model za binarnu klasifikaciju nebalansiranih skupova podataka zasnovan na vektorima podrške koji uzima u obzir i podatke koji nedostaju. Smieja et al. u [45] kreiraju ANN koja, bez imputacije, u prvom skrivenom sloju vrši modifikaciju izlaza tako da bude uključena očekivana vrednost MD. Ovakvi radovi su posebno značajni jer nadograđuju ideju o primeni ML za imputaciju i vode eventualno poptpuno novom pristupu posmatranju ovog problema, različitog od onog koji se obično podrazumeva i koji smo nasledeli iz statistike, a to je analiza kroz mehanizme i paterne nedostajanja.

4.4 Neuralne mreže kao metod imputacije u proceni CMR rizika

U nastavku je prikazana naša studija koju smo prikazali u [1] gde smo istraživali koju tačnost ostvaruje jednostruka imputacija uz pomoć ANN u odnosu na druge metode u okviru predikcije kardio-metaboličkog rizika .

Da bi se, uz pomoć mašinskog učenja, formirao prediktivni model koji koristi samo jednostavne, lako dostupne vrednosti kao što je [46] laboratorjski nalazi su neophodni samo u fazi pretporcesiranja kada se izračunavaju izlazne vrednosti budućeg modela. Ukoliko neke od takvih vrednosti nedostaju, uklanjanjem kompletnih podataka bi se značajno smanjio obučavajući skup i učenje bi bilo otežano ili čak nemoguće. Zato smo u [1] posmatrali kako bi se ANN ponašala kao metoda imputacije za ovakvu vrstu vrednosti.

Posmatrali smo laboratorijske vrednosti: HDL holesterol (HDL), LDL holesterol (LDL), ukupni holestrol (TCH), trigliceridi (TG) i glukoza (GLY). Sve one utiču na formiranje procene CMR rizika, ali se same ne koriste u formiranju finalnog prediktivnog modela.

Za potrebe imputacije smo koristili regresione ANN sa jednim skrivenim slojem i *resilient backpropagation* algoritmom za učenje [47, 48]. Za svaku promenljivu od interesa smo prvo pronašli broj skrivenih neurona ANN koji omogućava optimalnu predikciju ovih vrednosti na osnovu ostalih uključujući i ulazne i spoljašnje promenljive (laboratorijske vrednosti).

Simulacijama smo iz polaznog skupa uklanjali podatke, posebno za svaku laboratorijsku vrednost. Uklanjanje smo radili za različite količine MD (10%, 20%, 30% i 50%) i na osnovu sva tri mehanizma nedostajanja (MCAR, MAR, MNAR). Nad tako formiranim skupovima samo izvršili upoređivanje sledećih metoda: ANN, PMM (predictive mean matching), SLR (stohastic linear regression), RF (random forest) i MEAN (mean imputation). Radi se o jednostavnim jednostrukim metodama za imputaciju, koje su nam bile neophodne u početnoj fazi istraživanja MD kod spoljašnjih promenljivih pri razvoju prediktivnih modela. Sve metode su okviru statističke analize MD razmatrane detaljno u [12] i ostaloj literaturi. Upoređivanje metoda je izvršeno na osnovu tačnosti imputacije vrednosti kao i tačnosti naknadne klasifikacije samog rizika koja se dobija izračunavanjem na osnovu novih, dobijenih laboratorijskih vrednosti.

Detaljni rezultati, diskusija za svaku promenljivu i ograničenja su izneti u publikaciji [1] dok će u nastavku biti naveden opšti zaključak.

Tabela 4.1 prikazuje rezultate za tri metode koje su pokazale najbolje performanse, a to su: ANN, MEAN i RF. Prikazuje sumarni pregled slučajeva u kojima određena metoda ostvaruje najbolje rezultate gde jedan slučaj predstavlja kombinaciju promenljvie, mehanizma, količine i ocene performanse koja se odnosi na tačnost imputacije vrednosti i finalne klasifikacije rizika. Isti rezultati su prikazani grupisan po ovim kategorijama pa se. Na primer, može se videti da ANN ostvaruje najbolje rezultate ukupno u 133 posmatrana slučaja, 40 puta za MCAR mehanizam, 54 puta za MAR i 39 puta za MNAR mehanizam. Za MEAN i RF metode je, u zagradama, prikazan i broj slučajeva u kojima je ANN sledeća najbolja metoda.

Method	**Total number of winning cases**
ANN	133
MEAN	21 (18)
RF	26 (20)

by variable	**HDL**	**LDL**	**TCH**	**TG**	**GLY**
ANN	24	33	36	19	21
MEAN	8 (5)	0	0	1 (1)	12 (12)
RF	4 (1)	3 (3)	0	16 (13)	3 (3)

by mechanism	**MCAR**	**MAR**	**MNAR**
ANN	40	54	39
MEAN	12 (9)	4 (4)	5 (5)
RF	8 (6)	2 (2)	16 (12)

by volume	**10%**	**20%**	**30%**	**50%**
ANN	31	33	33	36
MEAN	4 (4)	4 (4)	6 (4)	7 (6)
RF	10 (9)	8 (5)	6 (5)	2 (1)

Tabela 4.1: Rezultati upoređivanja metoda imputacije kroz različite slučajeve MD koji prikazuju algoritme sa najboljim perfomansama grupisane prema promenljivoj, mehanizmu nedostajanja i količini MD. Za MEAN i RF je, u zagradam, prikazan broj slučajeva kada je ANN naredni najbolji metod. *Tabela je preuzeta iz publikacije [1]*

Iz rezultata se može videti da ANN ostvaruje najbolje rezultate najvećem broju slučajeva. Čak i u slučajevima gde neki drugi metod ispoljava bolje performanse, ANN predstavlja sledeći najbolji metod.

Zaključeno je da se ANN može iskoristiti kao metod imputacije spoljašnjih promenljivih prilikom formiranja prediktivnog modela za procenu kardio-metaboličkog rizika s obzirom da stabilno postiže najbolje rezultate u odnosu na ostale posmatrane metode prilikom promene mehanizama nastajanja ili količine nedostajućih podataka. Data je i napomena da istraživanje ne isključuje ostvarivanje dobrih rezultata primenom i nekog drugog algoritma ML i da, u tom slučaju, realizovana metodologija može predstavljati radni okvir za istraživanje drugih algoritama. Ovo je ujedno bio i prvi korak ka formiranju opšte metodologije koja je prikazana u nastavku disertacije.

5. Podaci koji nedostaju i razvoj prediktivnih modela mašinskog učenja

Prilikom razvoja prediktivnih modela mašinskog učenja, kao u bilo kom slučaju kada se istraživanje vrši nad podacima, pojavljuje se problem nedostajućih podataka. Međutim, prilikom primene nad podacima koji imaju MD se često, u fazi pretprocesiranja, pribegava uklanjanju podataka zašta smo već naveli da nije uvek potrebno, a može biti čak i škodljivo. Osim što istraživači koji se bave primenom mašinskog učenja nisu u tolikoj meri zainteresovani za ovaj problem, često ili u isto vreme nisu ni ni svesni potencijalnog uticaja na završne rezultate. Ovo se pogotovo dešava kod veće količine podataka gde ta količina može da utiče na pretpostavku da uklanjanje neće biti od značaja za dalju izgradnju modela. Iako nekada uklanjanje može biti opravdano, kod manjih količina MD i uz MCAR mehanizma na primer, ta odluka o uklanjanju bi trebala da bude opravdano doneta.

Drugi problem koji se javlja u domenu primene ML je kad istraživač odluči da uradi imputaciju, ali nekom ad-hoc izabranom metodom. Na primer, koristeći podrazumevani metod imputacije koji neke biblioteke za realizaciju ML algoritama nude u sklopu funkcija za pretprocesiranje podataka. Recimo, funkcija *preProcess* bilioteke *caret* [49] koja se često koristi u R okruženju za razvoj prediktivnih modela nudi mogućnost izbora dva tretmana nedostajućih podataka: izbaciti ili izvršiti imputaciju metodom k-najbližih suseda. Ovaj pristup, izbor predefinisane metode za imputaciju, takođe u nekim slučajevim može biti uspešan, ali bi, u opštem slučaju i ovu odluku trebalo opravdati.

Ovo sve znači da bi istraživač, u toku pretprocesiranja, trebalo da:

1. bude svestan problema MD i njegovih posledica odnosno da poznaje osnovnu teoriju MD,
2. utvrdi strukturu MD,
3. analizira veze među prisutnim i nedostajućim podacima i utvrdi paterne i mehanizme nedostajanja,
4. utvrdi da li uklanjanje opravdano,
5. ukoliko uklanjanje nije opravdano da dalje istraži metode koje mogu služiti za imputaciju, utvrdi zadovoljenje njihovih pretpostavki i odabere odgovarajući mehanizam imputacije.

Ako čak i pretpostavimo da je prvi korak zadovoljen, ceo dalji proces može biti veoma složen. Većini stručnjaka koji nisu iz oblasti statistike ili koji se nisu susretali ranije sa ovim problemom će poslednji koraci biti čak i nepoznati. Pored potrebne ekspertize ovi procesi zahtevaju i dodatno vreme potrebno za implementaciju. Sve to zajedno dalje može uticati na motivaciju da se uopšte pristupi navedenim procesima. Zato je jedan od ciljeva ovog istraživanja bio da se utvrdi metodologija koja bi omogućila razvoj modela uz dodavanje analize uticaja MD na finalni model, jednostavna za implementaciju (sa strane inženjera) i bez potrebe za postojanjem prethodne analize i ekspertize o MD. Ukoliko bi se utvrdilo da neki metod imputacije nije u značajnoj meri osetljiv na vrstu i količinu MD to bi omogućilo realizaciju automatske obrade MD što bi nas dalje približilo automatizaciji celokupnog procesa pretprocesiranja što se često navodi i kao jedan od najvažnijih zadataka razvoja mašinskog učenja [50].

5.1 Značaj razmatranja problema nedostajućih podataka

Problem podataka koji nedostaju se može pojaviti u različitim fazama razvoja i primene prediktivnog modela.

Prvo, podaci nad kojima vršimo konstrukciju modela i koji će služiti za treniranje i testiranje finalnih modela mogu imati MD. U tom slučaju struktura MD je poznata, ali bez dodatne analize ne možemo, u opštem slučaju, utvrditi zbog čega su ti podaci neodstupni i u kakvoj su vezi sa ostalim, postojećim podacima. Ovo i jeste najčešći slučaj kada se bavimo problemom MD jer smo tada prinuđeni da razmišljamo o odgovoru na pitanje: šta raditi sa MD? Upravo i ovde, dolazi do grešaka, u praktičnoj primeni ML metoda, koje su već navedene.

Drugo, podaci koje koristimo mogu biti kompletni, ali možemo imati osnovane pretpostavke da u nekim drugim instancama skupova podataka koje se potencijalno mogu koristiti za izgradnju modela mogu postojati MD. Ova situacija može odgovarati bilo kom istraživanju koje ne publikuje podatke, već samo predloge modela koje dalje stručna i naučna zajednica koriste za implementaciju nad sopstvenim podacima. U tom slučaju bi, prilikom izgradnje i publikacije rezulata, trebalo izvršiti određenu analizu ponašanja modela pod uslovima kad postoje MD.

Slično se dešava kada smo mi sami u ulozi onih koji primenjuju saznanja o nekom postojećem modelu. Često je za određeni problem poznat dobar prediktivni model. Međutim, ukoliko mi pokušamo da ga implementiramo nad drugim skupom podataka iste strukture koji sadrži MD može se pojaviti problem. Ukoliko je originalni model bio razvijan nad kompletnim skupom ili čak nad nekim koji je tretirao MD na nama nepoznat način, može se doći do različitih rezultata. U ovim slučajevima se može primeniti upoređivanje polaznih skupova podataka, koliko je to moguće, ali se to obično svodi na analizu osnovnih statistika jer su to često i jedine informacije koje se objavljuju uz određeni model. MD, kao potencijalni uzrok različitosti skupova, se retko i razmatra.

Treba spomenuti i slučaj kada možemo naslutiti, ili se s pravom zapitati, da li je uopšte skup nad kojim razvijamo model zaista kompletan skup. Ova situacija se javlja kad god podatke dobijamo od treće strane bez dodatne informacije o tome da li su i kako uklonjeni eventualni MD.

Na kraju, kad imamo gotov model i kad se on koristi u produkciji svako novo posmatranje koje zahteva predikciju može imati neke podatke koji nedostaju.

Zajedničko za sve navedene probleme i što doprinosi ukupnoj problematici je i to što je izveštavanje o postojanju i tretmanu MD često zanemaren deo u naučnim radovima koji se bave prikazom nekog novog ili boljeg modela u domenu razvoja prediktivnih modela mašinskog učenja nad određenim skupovima podataka. Ako i postoji, obično se radi o osnovnim informacijama: koliko je bilo vrednosti koje nedostaju i, eventualno, koji tretman MD je primenjen. Obično ne postoji detaljna analiza potproblema koji se odnosi na MD, a razlozi za izbor metoda tretmana MD veoma često nisu navedeni.

Sve prethodno navedene situacije su slične, a razlikuju se, suštinski, u jednoj stvari: da li je poznata struktura MD? U tom smislu razlikujemo tri slučaja:

1. Podaci nisu kompletni, čime nam je poznata stuktura MD,
2. Podaci su kompletni i poznata nam je struktura MD koji se mogu pojaviti,

3. Podaci su kompletni i nije nam poznata struktura MD koji se mogu pojaviti.

Kao što je navedeno, poznavanje strukture nije dovoljno da se sa sigurnošću može utvrditi da li se MD mogu zanemariti u procesu razvoja modela. Zato bi, u datim slučajevima, bilo od koristi imati jednu fazu razvoja gde bi se pažnja posvetila tretmanu MD. S obzirom da u domenu razvoja modela mašinskog učenja, i o čemu je već bilo reči, uglavnom ne postoji ekspertiza za ovaj proces, a neretko i motivacija za njegovo izvršavanje], onda bi određeni automatski procesi mogli doprineti lakšem usvajanju ovakve, nove, faze pretporecesiranja. U narednom poglavlju će upravo biti iznet predlog jedne takve metodologije.

5.2 Metodologija za analizu uticaja imputacija pri razvoju prediktivnih modela

U ovom poglavlju je predstavljena metodologiju koja se može iskoristiti za analizu uticaja nedostajućih podataka i njihovih imputacija pri razvoju prediktivnih modela za slučajeve kad polazni skupovi imaju ili potencijalno mogu imati podatke koji nedostaju. Predstavljena metodologija je usklađena sa domenom u kom se očekuje njena primena, a to je razvoj i primena prediktivnih modela zasnovanih na mašinskom učenju.

U nastavku će biti definisana dva algoritma:

- M1: Algoritam za simulaciju uticaja imputacije pri razvoju prediktivnog modela u prisustvu nedostajućih podataja (struktura MD je poznata),
- M2: Algoritam za simulaciju uticaja imputacije pri razvoju prediktivnog modela uz potencijalno prisustvo MD (struktura MD nije poznata).

U prethodnom poglavlju su navedene različite potrebe kada analiza uticaja imputacije može biti od koristi. Zavisno od slučaja korišćenja se može primeniti jedan od predloženih opštih algoritama. Prvi algorirtam je namenjen, prvenstveno, za pronalaženje i konstrukciju finalnog modela, dok je drugi namenjen za analizu različitih imputacija kod poznatog modela. Međutim, ovo ne mora biti pravilo. Ako polazni skup ima značajniju količinu MD, a konačni cilj je pronalaženje modela nad takvim skupom može se koristiti prvi algoritam. S druge strane, ako polazni skup ima veoma malu količinu MD ona se može ukloniti iz polaznog skupa, ali se kasnije iskoristiti kao informacija za drugi algoritam s obzirom da prisutnost nedostajnja često ukazuje na potencijalnu novu pojavu sličnog nedostajanja.

Algoritmi su dati u obliku pseudokoda i to njihovi najznačajniji strukturni delovi čime se obezbeđuje opštost. Nakon definicije svakog algoritama data su objašnjenja i uputstva za njihovu implementaciju.

Navedeni algoritmi su u parametrizovani u značajnoj meri. Zbog toga se može pojaviti i veća složenost procesa pa, shodno tome, treba voditi računa o pametnom izboru polaznih parametara. Ovde se, pre svega, misli na izbor finalnih prediktivnih modela i modela koji se koriste za imputaciju. Ukoliko obe vrste modela nastaju kao rezultat primene algoritama mašisnkog učenja njihovo obučavanje samo po sebi može imati veliku složenost. Zato su psoebno date napomene na svim mestima gde donošenje odluka može biti značajno za usložnjavanje ukupnog procesa. Navedene napomene takođe treba smatrati za deo predložene metodologije.

Dobar princip kojim se ovde treba voditi je princip Okamove oštrice [51] i što je moguće češće birati jednostavnije postavke u kontekstima gde je osnovano pretpostaviti da se rezultati ne bi značajno poboljšali usložnjavanjem procesa. Na primer, ako smo već utvrdili da nad sličnim podacima ANN preciznije imputira podatke nego LR onda nema potrebe upoređivati ove dve metode već je dovoljno koristiti ANN. S druge strane, ako nam ANN sa složenijom strukturom ne donosi velike razlike u tačnosti imputacije onda je dovoljno koristiti ANN sa jednostavnijim hiperparametrima s obzirom da se vreme obučavanja može značajno skratiti.

Pre prikaza metodologija biće prikazan osvrt na česte pojmove i tehnike koji se koriste u procesima analize imputacije MD i razvoju prediktivnih modela kao sugestije koje se mogu iskoristiti prilikom primene datih algoritama.

5.2.1 Amputacija podataka

Za potrebe istraživanja je često potrebno izvršiti simulaciju MD odnosno namenski ukloniti podatke iz polaznog skupa kako bi se mogao ispitati uticaj nedostajanja na formiranje modela. Ukoliko se ovakava simulacija izvršava neplanski i nestruktuirano teško bi se mogla izvršiti struktuirana i smislena analiza koja se odnosi na složenije mehanizmi i strukture nedostajanja koje odgovaraju realnim situacijama. Uklanjanje podataka iz skupa kompletnih podatka se, nasuprot imputaciji, još naziva i *amputacija* (eng. amputation)

U eksperimentalnom delu istraživanja ove disertacije, je intenzivno korišćen proces amputacije. Ovaj proces je zasnovan na primeni funkcije *ampute* paketa mice [52, 53] u R programskom okruženju.

Osnovne strukture uz pomoć kojih se vrši modeliranje amputacija odnosno uklanjana MD prilikom različitih scenarija su:

- **skup paterna** u obliku matrice čije vrste predstavljaju paterne nedostajanja kao što je ranije definisano,
- **skup težina** u obliku matrice čiji redovi predstavljaju koeficijente koji se koriste za izračunavanje težinskog skora za odgovarajući patern i time modeluju zavisnost među promenljivama i mehanizme nedostajanja,
- **skup pravila o distribuciji** u obliku matrice čiji redovi predstavljaju opis na koji način će se MD pojavljivati u odnosu na rezultate težinske funkcije,
- **vektor učestalosti** čija svaka komponenta određuje procenat pojavljivanja odgovarajućeg paterna u ukupnom skupu paterna,
- **mehanizam nedostajanja** sa napomenom da se na mehanizam može uticati i preko skupa zavisnosti odnosno težina,
- **količina** MD podataka sa značenjem koje je ranije navedeno.

Iako je ovaj skup struktura formiran za potrebe realizacije specifične funkcije u jednom programskom jeziku, sam postupak i ideja modelovanja amputacija je opšteg karaktera i može se primeniti u bilo kom programskom okruženju.

Pomoću navedenih struktura podataka i njihovih vrednosti mogu se modelirati najrazličitiji scenariji pod kojima se javljaju MD. One su korišćene i u realizaciji eksperimenata u našem radu [1], a i u studiji slučaja prikazanoj u ovoj disertaciji u 6.4.2.

5.2.2 Performanse prediktivnih modela

Prilikom izgradnje i analize ML modela neophodno je meriti njihove performanse. Ovo obično podrazumeva evaluaciju ponašanja modela prilikom predikcija nad novim test skupovima podataka, ali i prilikom izbora hiperparametara i kod upoređivanja različitih modela.

Performanse modela se kvantifikuju na različite načine. Koje mere će biti korišćene zavisi od tipa problema (regresija, klasifikacija), ali i od ciljeva istraživanja. U nastavku je navedeno nekoliko standardnih mera koje se koriste u domenu ML i koje su korišćene u eksperimentalnom delu istraživanju.

Ako sa Y označimo izlazne vrednosti polaznog skupa, sa $\hat{Y}$ skup predikcija dobijenih modelom, a sa n broj opservacija u polaznom skupu navedene mere

predstavlja grešku modela nad ulaznim podacima. Ekvivalentne formule se koriste i za evaluaciju nad spoljašnjim test skupovima.

Mere koje su korišćene za regresione probleme su sledeće.

MSE

Mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5.1}$$

Nastala od ukupne apsolutne greške kako bi se izbeglo izračunavanje apsolutne vrednosti.

RMSE

Root mean squared error

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{5.2}$$

Nastala od MSE radi lakše interpretacije jer su jedinice iste kao kod ulaznih podataka.

MAPE

Mean absolute percentage error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| 100\% \tag{5.3}$$

Objašnjava procentualnu grešku i iako ima određene mane [54] ima intuitivnu interpretaciju pa je zato uključena prilikom upoređivanja imputacija.

U problemima klasifikacije se obično uzimaju mere nastale iz vrednosti koje se dobijaju formiranjem *matrice konfuzije*. Matrica konfuzije predstavlja pregled ispravno i pogrešno klasifikovanih podataka na osnovu koje se izvode određene vrednosti i zaključci.

Primer matrice konfuzije za binarnu klasifikaciju je prikazan u tabeli 5.1. Sa 0 i 1 su označene klase u binarnom klasifikacionom problemu. One često pred-

		Predikcija	
		1	0
Stvarni podaci	1	TP	FN
	0	FP	TN

Tabela 5.1: Matrica konfuzije za binarni klasifikacioni problem

stavljaju informaciju o tome da li neki podatak ima (1) ili nema neku osobinu (0). Na primer, u medicinskoj dijagnostici 0 i 1 bi označavale postojanje određee dijagnoze, stanja ili sindroma. Otuda i najčešće korišćene oznake vrednosti u matrici konfuzije: TP (stvarno pozitivni, eng. true positives), TN (stvarno negativni, eng. true negatives), FP (lažno pozitivni, eng. false positives) i FN (lažno negativni, eng. false negatives). U svakom konkretnom slučaju se u matrici navodi ostvaren broj navedenih slučajeva. Koristeći dobijene vrednosti formiraju se pravila za određene mere perfomansi klasifikacionih problema.

Accuracy (Tačnost)

$$Acc = \frac{TP + TN}{P + N} \tag{5.4}$$

Predstavlja odnos tačnih predviđanja i ukupnog broja predviđanja. Sa P je predstavljen realan broj pozitivnih (1), a sa N realan broj negativnih slučajeva (0). Iako prikazuje opštu tačnost, za klasifikacione probleme može biti neinformativna mera pogotovo ako skup nije balansiran u odnosu na postojeće klase.

Precision (Preciznost)

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

Meri da li ima puno onih koji su pogrešno klasifikovani kao pozitivni jer meri odnos stvarno pozitivnih i ukupan broj onih koji su klasifikovani kao pozitivni. Mala preciznost ukazuje da ima puno slučajeva koji su pogrešno klasifikovani kao pozitivni. Drugi, često korišćen naziv je i PPV (eng. positive predicitive value).

Recall (Odziv)

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{5.6}$$

Obrnuto od preciznost, odziv meri koliko ima onih koji su pogrešno klasifikovani kao negativni jer meri odnos stvarno pozitvnih i realan broj pozitivnih. Mali odziv ukazuje da ima puno onih koji pogrešno nisu prepoznati kao pozitivni. Naziva se još i *Sensitivity*.

Specificity (Specifičnost)

$$Specificity = \frac{TN}{N} \tag{5.7}$$

Meri tačnost negativno predviđenih slučajeva jer predstavlja odnos ispravno predviđenih i realnog broja negativnih slučajeva.

F1 score

$$F1 = 2\frac{Precision * Recall}{Precison + Recall} \tag{5.8}$$

Koristi se kao harmonijska vrednost preciznost i odziva. Idealno je da obe vrednosti budu visoke (blizu 1) ali se često dešava nagodba (eng. trade-off) ove dve vrednosti u smislu da poboljšanje jedne vrednosti može dovesti do pogoršanja druge. Koliko je ovo bitno i u kom smeru (veća ili manja preciznost ili odziv) zavisi, pre svega, od vrste problema. Navedena mera predstavlja balansirani F-score u kom obe vrednosti imaju isti značaj. Formula se može proširiti tako da omogućava povećavanje značaja jedne od ove dve vrednosti.

5.2.3 Simulacije u okviru mašinskog učenja

Metodologija razvoja modela mašinskog učenja, bez obzira na izbor i implementaciju algoritama, obično podrazumeva neku vrstu simulacije koja omogućava pronalaženje što ispravnijeg modela. Pod ispravnim modelom podrazumevamo model koji izvršava zadatak sa određenom tačnošću i to ne samo na obučavajućem skupu već i na odvojenim test skupovima, a potencijalno i idealno isto tako i na novim slučajevima koji se kasnije, dok se model primenjuje pojavljuju.

Ove simulacije se obično izvršavaju u 2 navrata:

- prilikom podešavanja hiperparametara,
- prilikom testiranja modela.

Simulacije se mogu izvršavati pojedinačno, kao jedinstveni procesi, a mogu postojati i dodatne simulacije specifične za određene probleme. Na primer, ukoliko se zbog određenih sigurnosnih ograničenja ne smeju koristiti realni podaci mogu se simulirati posebni podaci koji preslikavaju odnose iz realnog skupa. Takođe, male baze podataka je ponekad neophodno multiplicirati bez ponavljanja podataka. Važan slučaj, koji se odnosi i na ovaj rad, su simulacije koje omogućavaju ispitivanje posledica ili validaciju nekih odluka i akcija u cilju postizanja boljih konkretnih, finalnih prediktivnih modela.

U ovoj disertaciji simulacije su korišćene kao sredstva validacije:

- prilikom pretrage prostora i utvrđivanja hiperparametara ML algoritama,
- prilikom obučavanja i testiranja ML modela,
- prilikom upoređivanja performansi nad različitim skupovima [55, 56].

Korišćene su sledeće metode:

- bootstrap 635 with replacement resampling ([57]),
- MonteCarlo unakrsna validacija (eng. Monte Carlo or random leave group out cross-validation),
- k-slojna unakrsna validacija (eng. k-fold cross-vaidation, u daljem tekstu kCV),
- k-slojna unakrsna validacija sa ponavljanjem (eng. repeated k-fold cross-validation).

U pojedinim slučajevima, metode su korišćene zasebno, a u drugim i kao ugnježdene metode gde se nad svim podskupovima podele jedne metode vrši nova validacija kroz novu podelu skupova.

5.3 M1: Uticaj imputacije pri razvoju prediktivnog modela u prisustvu nedostajućih podataka

Nakon tehničkih pojmova koji se koriste u objašnjenjima algoritama u ovom poglavlju je opisan prvi algoritam za analizu uticaja uticaja MD pri razvoju prediktivnih modela. Ovaj algoritam se odnosi na razvijanje modela nad skupom koji sadrži MD odnosno u slučajevima kada je struktura MD poznata.

Opisani algoritam predstavlja opšti formalni postupak za testiranje jednog ili više modela imputacije prilikom razvoja finalnog prediktvnog ML modela. Cilj

postupka je utvrditi da li je i u kojoj meri finalni model osetljiv na promene strukture i mehanizme nedostajanja nekih vrednosti. Ovo se utvrđuje upoređivanjem performansi finalnih modela nad skupom koji sadrži kompletne podatke (što odgovara uklanjanju MD) i nad skupovima koji su dobijeni imputacijom jedne ili više metoda.

5.3.1 Algoritam

Listing 5.1: Algoritam M1 za analizu uticaja imputacija nad podacima koji sadrže MD prilikom razvoja prediktivnog modela mašinskog učenja

```
# PREPARATION
#########################################################################
data <- read data with MD: X, Y

F <- define final models: F_1, F_2, ..., F_fn
P <- gather patterns from data: p_1, p_2, ..., p_pn
IM <- define imputation methods: im_1, im_2, ..., im_imn
SP <- define sim params

data.cc <- get complete cases
foreach f in F
   fit and cross validate model on data.cc
   cc.results <- update
report cc.results


# SIMULATION
#########################################################################

use data
foreach im in IM
   foreach p in P
      imp.sets <- make sets: data.imp.input, data.imp.output, data.imp.test
      p.data.imputed <- impute for p on data

   data.imputed <- merge p.data.imputed
   data.imputed <- recalcuate values if needed

   TP <- make partitions for final model test on data.imputed:
            (train_1, test_1), (train_2, test_2), ..., (train_tpn, test_tpn)

   foreach (train, test) sets in TP
      foreach f in F
         fit and cross validate model f
         partition.f.results <- update
      partition.results <- gather all partition.f.results
```

```
37
38    finalmodel.results <- gather all partition.results
39
40   im.results <- gather all finalmodel results
41
42   report how volume, number of vars and md mechanism
43   affect final models for im method
44
45 final.results <- gather all im.results
46 compare performance for all cases, including complete-case
47 choose imputation and final method
```

5.3.2 Uputstva

Prilikom navođenja napomena u ugulastim zagradama je navedena linija algoritma, prikazanog u 5.1, na koju se napomena odnosi.

Priprema - učitavanje podataka [3]

Skupovi $X = (x_1, x_2, ..., x_{xn})$ i $Y = (y_1, y_2, ..., y_{yn})$ predstavljaju ulazni i izlazni skup podataka. Čest je, ali ne i obavezan, slučaj da je $yn = 1$.

Ovde treba obratiti pažnju na dve važne stvari prilikom realizacije algoritma.

Prvo, pod ulaznim skupom X smatramo sve dostupne podatke, pa i one koji potencijalno neće učestvovati ni u jednom finalnom modelu s obzirom da se oni mogu iskoristi za imputacije. Ovo implicira da proces analize uticaja MD ne treba vršiti nakon procesa konačnog izbora podskupa promenljivih koje učestvuju u izgradnji finalnog prediktivnog modela.

Dalje, važno je iznova skrenuti pažnju da se u ovom momentu ne smeju vršiti nikakva pretprocesiranja koja se odnose na standardizaciju ili normalizaciju podataka. Ukoliko je planirano korišćenje neke od navedenih procedura ona se sme izvršiti tek u momentu kad se formiraju skupovi neophodni za realizaciju modela (treniranje, validacija, testiranje) i to na skupovima za obučavanje kako ne bi došlo do curenja podataka (eng. data leak).

Priprema - Finalni modeli [5]

Skup $F = \{f_1, f_2, ..., f_{fn}\}$ je skup svih finalnih prediktivnih ML modela koji se razmatraju u istraživanju i predstavlja podskup svih mogućih funkcija zavisnosti X i Y.

U ovom koraku je potrebno definisati koji finalni prediktivni modeli se istražuju. Ukoliko se radi o istraživanju u kom nam je unapred poznat model koji se koristi ovaj skup ima samo jedan elemenat. Ukoliko se nalazimo na početku novog istraživanja i tek okrivamo potencijalni model onda ovaj skup može imati više modela. U tom slučaju, ukoliko je neophodno preciznije praćenje analize korak po korak, algoritam se može razdvojiti na zasebno izvršavanje postupka za svaki finalni model.

Priprema - Paterni [6]

Skup $P = \{p_1, p_2, ..., p_{pn}\}$ predstavlja skup paterna MD koji se javljaju u polaznom skupu.

Neophodno je prepoznati i definisati paterne nedostajanja iz ulaznih podataka na osnovu redova koji sadrže nepoznatu vrednost za bar jedan ulazni parametar. Ovom prilikom se može evidentirati i učestalost pojavljivanja pojedinačnih paterna koja se posle može iskoristiti za dodatnu analizu MD ili u prikazu konačnih rezultata.

Prilikom implementacije, za programsku realizaciju paterna, se mogu iskoristiti strukture slične onima opisanim u poglavlju 5.2.1.

Priprema - metode imputacije [7]

Skup $IM = \{im_1, im_2, ..., im_{imn}\}$ predstavlja skup metoda koje će se koristiti za imputaciju.

Za metod imputacije se može koristi bilo koji metod pa i metode ML. S obzirom da je jedan od ciljeva ovog postupka da isključimo bavljenje dodatnom analizom samih MD, algoritmi ML zapravo i jesu dobri kandidati za imputaciju.

Postupak izbora metode imputacije je u direktnoj vezi sa drugim postupcima koje činimo u okviru pretprocesiranja, kao što su analiza korelacija ili linearnih zavisnosti među promenljivama odnosno proces izbora samih promenljivih koje će se koristiti kao prediktori. Saznanja iz takve analize mogu biti od koristi za izbor metoda imputacije [to opet ukazuje na to da proces analize MD ne treba vršiti na kraju već u sklopu ostalih aktivnosti faze pretprocesiranja.

Takođe, treba obratiti pažnju na specijalan slučaj kada se isti algoritam ML koristi za imputaciju i finalni model. Iako se i taj slučaj može testirati, može se desiti da se neće javiti značajnije razlike u rezultatima. Ovo nastaje zbog toga jer se prvo na osnovu jednog podskupa vrši imputacija nekih vrednosti, a onda se

slična procedura ponavlja nad proširenim, popunjenim podacima. Iako se, formalno gledajući, model za imputaciju i finalni model razlikuju, saznanja iz prvog modela predstavljaju podskup saznanja iz drugog jer su informacije iz prvog skupa delom sadržane i u drugom.

Na kraju, treba voditi računa o tome da izbor skupa IM zavisi od skupa P utoliko što je za svaki patern p_i potrebno definisati koji model imputacije se koristi za dati patern. Može se raditi o različitim algoritmima ML, ali i o istom ML algoritmu. Na primer, ANN se može koristiti kao metod imputacije za svaki patern s tim da će strukture tih ANN modela biti različite za svaki patern.

Priprema - Parametri simulacije[8]

Skup $SP = \{sp_1, sp_2, ..., sp_{spn}\}$ je skup parametara neophodnih za izvršavanje simulacija.

Nepohodno je u postupku koristiti neki metod unakrsne validacije i za njega je neophodno definisati parametre. Na primer, za kCV sa ponavljanjem bi to bio broj slojeva k i broj ponavljanja r dok za MC validaciju dodatni parametri nisu neophodni.

Priprema - Kompletni podaci[10]

Kao referencu, krierati i testirati finalne modele na podskupu podataka koji nemaju MD ukoliko je to moguće tj ukoliko postoji određeni deo skupa koji sadrži kompletne podatke. Ovo posebno ima smisla uraditi ako polazni skup ima manju ukupnu količinu MD. Isto tako, u slučaju da jedna promenljiva ima puno nedostajanja intuitivna bi pretpostavka bila da se ona isključi iz skupa ulaznih podataka. Tada je korisno uraditi imputaciju tih vrednosti jer se može ispostaviti da je takva pretpostavka pogrešna.

Simulacija - priprema skupova[23]

Skup $imp.sets = \{data.imp.input, data.imp.output, data.imp.test\}$ je skup od tri elementa i predstavlja jednu particiju skupa gde je polazni skup podeljen na ulazne vrednosti, izlazne vrednosti i test vrednosti.

Da bi se izvršila imputacija koja podrazumeva korišćenje nekog algoritma ML njegova konstrukcija podrazumeva postojanje ulaznih i izlaznih podataka. Ovde treba voditi računa koji podaci ulaze u skupove za imputaciju ukoliko skup sadrži neke zavisne promenljive koje se dobijaju izračunavanjem na osnovu vrednosti drugih promenljivih. U ovom smislu, test skup ne predstavlja test skup za testi-

ranje finalnog modela već test skup za imputaciju pa predstavlja upravo podatke za koje je potrebno izvršiti imputaciju. Ukoliko se koristi neka jednostavna metoda imputacije, na primer imputacija nekim parametrom kao što je srednja vrednost, onda ovi skupovi nisu nepohodni.

Simulacija - imputacija[24]

U algoritmu, kao je definisan, je razdvojena petlja za imputaciju svake promenljive posebno. Podsetimo se da jedan patern predstavlja informaciju o tome koje promenljive nedostaju u redovima kojima patern odgovara. To znači da kad posmatramo jednu promenljivu, ostale koje su prisutne kao deo MD paterna obično ne mogu učestvovati u obučavanju.

Ranije smo naveli da svaki patern ima svoj model imputacije. Pošto patern može i najčešće predstavlja nedostajanje više vrednosti, model za imputaciju paterna može podrazumevati jedan model koji kao izlaz ima više vrednosti, ukoliko to model dozvoljava, ili može predstavljati skup modela gde svaki vrši imputaciju jedne promenljive. U tom slučaju je potrebno izvršiti posebne imputacije i na kraju ih sve spojiti u jedan skup na sledeći način:

```
foreach var in p
   data.train.imp <- impute for var
data.train.imp <- merge data.train.imputed
```

Ovaj izmena algoritma je zahtevnija za izvršavanje, ali može biti korisna ukoliko se analiza vrši korak po korak uz praćenje nastalih rezultata. Takođe, ovaj način podrazumeva velik broj obučavanja istih modela nad sličnim skupovima pa bi se u tom pravcu mogla izvršiti optimizacija.

Simulacija - kreiranje popunjenog skupa[26]

Svaki patern može izvršiti popunjavanje skupa za redove koji su mu odgovarajući. Međutim, moguće je i stalno vršiti izmenu početnog skupa. Time, sa jedne strane, dobijamo više podataka za amputaciju, ali sa druge strane taj skup u sebi može nositi značajne greške prouzrokvane pre svega nesigurnošću koje donose sami MD zajedno sa greškama same imputacije. Zato je, u osnovnoj definiciji algoritma, ostavljena procedura kad se svaka imputacija vrši uz pomoć originalnog skupa. Svakako, predloženo stalno menjanje ulaznog skupa se jednostavno može implementirati.

Bez obzira na način implementacije popunjavanja, nakon svih imputacija je potrebno izvršiti ponovna izračunavanja ili procene vrednosti zavisnih promenljivih ukoliko takve promenljive postoje.

Simulacija - particija podataka[29]

Skup $TP = \{\{train_1, test_1\}, \{train_2, test_2\}, ...\{train_{tpn}, test_{tpn}\}\}$ predstavlja particiju podataka prilikom validacije finalnog prediktivnog modela.

Ovde treba skrenuti pažnju na važnu stvar. Predloženim algoritmom su prvo izvršene imputacije celog skupa, a nakon toga je izvršeno obučavanje modela. Ukoliko polazni skup podataka sadrži neke MD to može biti indikacija da će i u budućnosti novi primeri koji budu zahtevali predikciju takođe imati MD. Postavlja se pitanje: da li particija nakon imputacije ima istu vrstu problema kao i standardizacija podataka pre particije? Odnosno, da li je neophodno imati izdvojene (eng. hold-out) test primere koji imaju MD? U vezi sa tim je i predikcija novih slučajeva koji imaju MD. Da bi se ovo rešilo, algoritam se može modifikovati tako da se čuvaju obučeni modeli imputacije nad trening podacima koji bi se kasnije koristili za imputaciju test skupova odnosno, u primeni, novih slučajeva. Time bi se i particija mogla izmestiti van petlje simulacije.

Navedena definicija algoritma, ipak zadržava dati oblik radi jednostavnijeg praćenja međurezultata i analize, koja je zapravo osnovni cilj algoritma. Takođe, ako se algoritam koristi u preliminarnim stadijumima ili samo i isključivo kao alat za analizu uticija imputacije, u navedenom obliku jeste dovoljan. Ukoliko je neophodno, navedena modifikacija se jednostavno može implementirati.

Što se tiče realizacije same particije, izdvajanje skupova se može izvršiti na bilo koji način. Opšte smernice su da se koristi kCv sa ponavljanjem i da se izbegavaju bootstrap metode za spoljne validacije finalnog modela.

Upoređivanje modela [44]

Za upoređivanje performansi finalnih prediktvnih modela se mogu koristiti bilo koje mere te namene. Izbor zavisi od: vrste problema, krajnih ciljeva i preferencija istraživača. U prethodnom odeljku 5.2.2 su navedene neke od njih koje se mogu iskoristiti u ovoj fazi, a koje su i korišćene u okviru ovog istraživanja.

5.3.3 Opšte napomene

Kao što smo naglasili na početku poglavlja, algoritam koji je predstavljen je opšteg tipa i može biti vrlo složen i računski zahtevan ukoliko se primeni za analizu koja podrazumeva razmatranje većeg broja različitih algoritama. Preporuka je unapred razmotriti skupove finalnih prediktivnih modela i modela imputacije. Da bi se ovo postiglo može se unapred, brzim testovima, izvršiti upoređivanje podskupa modela. Ako neki od njih na sličan način opisuju zadati problem izbarati jedan za koji će se dalje uraditi detaljna analiza. Zato je savet postaviti inicijalne skupove parametara na one sa što manje članova, po mogućustvu 1 do 2.

Napominjemo, cilj ovog algoritma nije da se poptuno precizno nađe najbolje moguće rešenje za imputaciju i konačni finalni model već predstavlja okvir pomoću kog se, analizom i zaključcima dobijenim iz njega, može formirati bolji model koji bi eventualno promakao istraživaču zbog ignorisanja MD. Počev od česte situacije da se nepotrebno odustane od neke promenljive i njenih dostupnih vrednosti. Kada se, nakon ovog postupka, izabere metod imputacije i finalni model onda se mogu izvršiti sva navedena podešavanja i optimizacije vezane za konkretne modele pa se u nastavku mogu formirati finalni prediktivni modeli boljih performansi.

Hiperparametri

U vezi sa prethodno navedenim je i proces pronalaženja i podešavanja hiperparametara koji su neophodni za realizaciju modela mašisnkog učenja koji se koriste i za imputacije i za finalne modele. Svakako se na odgovarajućim mestima može izvršiti poseban proces pretrage prostora hiperparametara. U tom slučaju, kad bi takva pretraga bila blizu imputacija ona bi se izvršavala puno puta nad sličnim podacima i na taj način bi odgovarala MonteCarlo ili bootstrap simulacijama za optimizaciju parametara.

Ukoliko se ne radi o većoj količini MD, a skladu sa ciljem da se postupak održi što jednostavnijim i dovoljno informativnim, onda se postupak pretrage prostora hiperparametara može izdvojiti i izvesti i u fazi pripreme. U tom slučaju bi se, na početku i pre simulacije, izvršilo sledeće:

```
foreach f in F
   cv for hyperparameters search for f

foreach im in IM
   foreach p in all P
      cv for hyperparameters search for im
```

Ukoliko je to moguće, skupovi za testiranje mogu biti srazmerni veličini koja odgovara količini MD koja se ispituje.

Ponovljivost

Radi kontrolisanog ponovnog pokretanja iste simulacije neophodno je pažljivo postaviti parametre koji se koriste prilikom simulacije slučajnih procesa, a to obično podrazumeva postavljanje inicijalne (eng. seed) vrednosti koju algoritmi za generisanje nizova pseudo-slučajnih vrednosti koriste. Minimalno kad se o tome treba povesti računa je prilikom kreiranja particija. S obzirom da se krieranje particija vrši unutar petlje, *seed* treba postaviti direktno ispred kreiranja particija i to na neku vrednost koja je sama ponovljiva. Na primer, to može biti vrednost koja zavisi od indeksa petlje ili vrednost iz nekog unapred definisanog niza vrednosti.

5.4 M2: Uticaj imputacije pri razvoju prediktivnog modela uz potencijalno prisustvo nedostajućih podataka

U nastavku je predstavljeni opšti formalni postupak simulacije MD za testiranje jednog ili više modela imputacije prilikom razvoja nekog finalnog prediktivnog modela, ovaj put nad kompletnim polaznim skupom. Cilj postupka je isti, utvrditi da li je i koliko finalni model osetljiv na promene strukture i mehanizama nedostajanja nekih vrednosti, ali je algoritam nešto složeniji s obzirom da nam je struktura MD nepoznata.

5.4.1 Algoritam

Listing 5.2: Algoritam M2 za analizu uticaja imputacija nad podacima sa potencijalnim MD prilikom razvoja prediktivnog modela mašinskog učenja

```
1 # PREPARATION
2 ######################################################################
3 data <- read data without MD: X, Y
4
5 F <- define final models: F_1, F_2, ..., F_fn
6 S <- define all MD scenarios: S1, S2, ..., S_sn
7 P <- define patterns for amputation: p_1, p_2, ..., p_pn
8 IM <- define imputation methods: im_1, im_2, ..., im_imn
9 SP <- define sim params: {amp.simNo, sp_1, sp_2, ..., sp_spn}
```

```
10
11
12 # SIMULATION
13 ##########################################################################
14
15 use data
16 TP <- make partitions for final model test on data
17         (train_1, test_1), (train_2, test_2), ..., (train_tpn, test_tpn)
18
19 # Original set
20 foreach f in F
21    fit and cross validate model using TP
22    startset.f.results <- update
23 startset.results <- gather all startset.f.results
24
25 # CC cases
26 forach s in S
27    foreach data.train and data.test set in TP
28       foreach i from 1 to simNO
29          data.train.amputed <- ampute using s
30          data.train.cc <- get complete cases from data.train.amputed
31          foreach f in F
32             fit and cross-validate f
33             cc.f.results <- update
34          cc.sim.results <- gather all cc.f.results
35    cc.s.results <- gather all cc.sim.results
36 cc.results <- gather all cc.s.results
37
38 # Imputations
39 foreach s in S
40    foreach im in IM
41       foreach data.train and data.test set in TP
42          foreach i from 1 to simNO
43             data.train.amputed <- ampute using s
44
45                foreach p in P
46                   imp.sets <- make sets:
47                   data.imp.input, data.imp.output, data.imp.test,
48                   (data.imp.actuals)
49
50                   p.data.imputed <- impute for p
51
52                data.imputed <- merge p.data.imputed
53                data.imputed <- recalcuate values if needed
54
55                foreach f in F
56                   fit and cross-validate f
57                   f.results <- update
58
59                sim.f.results <- gather all f.results
60             sim.results <- update
```

```
61          partition.results <- gather all sim.results
62        im.results <- gather all partiton.results
63     s.results <- gather all im.results
64 final.results <- gather all s.results
65
66 compare performances for all scenarios
```

5.4.2 Uputstva

U nastavku su navedena uputstva za korake koji su novi ili se razlikuju od prethodnog algoritma navedenih u 5.3.2, s tim da se smatra da tamo navedena uputstva i dalje važe. Kod svake napomene je u uglastim zagradama je naznačena linija algoritma prikazanog u 5.2.

Priprema - definisanje simulacija[6]

Skup $S = (S_1, S_2, ..., S_{sn})$ predstavlja skup scenarija koji zajedno predstavljaju jedan slučaj ispitivanja MD.

Kad imamo skup podataka koji već ima MD njihova struktura nedostajanja je poznata. Ovde, cilj je izvršiti simulaciju uticaja različitih scenarija gde jedan scenario predstavlja strukturu, mehanizam nastajanja i količinu MD. Na primer, cilj može biti analiza uticaj nedostajanja kod svake promenljive posebno ili analiza različite količine za isti mehanizam MD. Zbog toga je prvi korak analize definisanje scenarija koje je potrebno ispitati.

Iako će u pozadini simulacije postojati formalni mehanizmi nedostajanja korišćenjem ovog algoritma se definišu razmatrajući: koje količine MD su od interesa, kod kojih vrednosti se javljaju i da li su ta nedostajanja uslovljena nekim drugim vrednostima. Iako, teoretski, možemo formirati scenarije kao kombinacije varijacija svih promenljivih i vrednosti uticaja iz nekog skupa vrednosti slučajne promenljive, to je, u opštem slučaju, nepotrebno. Verovatno, većina tako generisanih scenarija ne bi odgovarala potencijalnim realnim situacijama kad postoje MD. Zato, u ovoj fazi, prilikom odabira scenarija, treba iskorisiti domensko znanje, poznate zavisnosti i korelacije među promenljivama iz prethodnih istraživanja ili prakse.

Priprema - Paterni[7]

Skup $P = \{p_1, p_2, ..., p_{pn}\}$ predstavlja skup paterna MD koji odgovara jednom ispitivanom scenariju.

Ovaj put, umesto prepoznavanja paterna iz podataka, je potrebno generisati paterne u skladu sa ciljevima istraživanja i kreirti podskupove za sve scenarije za koje se vrši analiza. Tom prilikom, ili koristeći neka prethodna znanja i pretpostavke ili proizvoljno, treba definisati različite paterne koji podrazumevaju: različite mehanizme (MCAR, MAR, MNAR), različitu količinu i različitu strukturu nedostajanja. Takođe je poželjno definisati i različite distribucije za nedostajnje kao i međusobni odnos učestalosti ponavljanja paterna.

Kao i za prethodni algoritam, preporuka za programsku realizaciju paterna su strukture opisane u poglavlju 5.2.1.

Priprema - Parametri simulacije[9]

Skup $SP = \{amp.simNo, sp_1, sp_2, ..., sp_{spn}\}$ predstavlja skup parametra neophodnih za realizaciju simulacija.

U ovom delu postupka se vrše simulacije koje veštački uvode MD tako da je, uz parametre unakrsnih validacija za obučavanje finalnih prediktivnih modela, potrebno definisati i paramatre potrebne za simulaciju MD.

Simulacija - particija podataka[16]

Što se tiče particije podataka, situacija je obrnuta u odnosu na prethodni algoritam. U ovom algoritmu se particija vrši na početku, ali se onda, iz istih razloga koji su navedeni u napomenama prethodnog algoritma , postavlja pitanje da li particiju treba izvršiti nakon amputacije podataka? S obzirom da je namena ovog algoritma analiza uticaja imputacije MD, a ne pronalaženje novog modela, osnovna definicija algoritma podrazumeva particiju pre simulacije kako bi se smanjila složenost procesa jer se na ovaj način mogu koristiti isti skupovi nad kojima je izvršena amputacija i za obučavanje finalnih modela i nad CC i nad popunjenim skupovima. Ukoliko je neophodno, CV petlja za particiju se može prebaciti neposredno pre obučavanja finalnog modela uz modifikaciju da se prethodno izvrši imputacija i u test podacima.

Simulacija - Početni model[20]

Radi reference, treba konstruisati modele i nad početnim skupom pre uvođenja MD odnosno izvršavanja simulacija amuptacija. Ovaj proces je, praktično, ekvivalentan postupku podešavanja hiperparametara i validaciji finalnih modela.

Simulacija - Kompletni podaci[26]

Pod kompletnim podacima (CC) se podrazumevaju podskupovi koji sadrže kompletne podatke. Ovi skupovi su istovetni za jedan scenario, bez obzira na model imputacije. Kako bi se izbegla nepotrebna idenična obučavanja finalnih modela ova faza se može izvršiti kao posebna simulacija. Važna pretpostavka u ovom slučaju je da su simulacije obavezno kontrolisano ponovljive i da su skupovi amputacija u obe simulacije identični. Ukoliko to nije slučaj rezultati neće biti uporedivi.

Simulacija - Amputacija[29, 43]

Kako bi se testirali različiti scenariji neophodno je obezbediti dobar metod amputacije. Za implementaciju ovog postupka se mogu iskoristiti strukture predložene u poglavlju 5.2.1.

Simulacija - Imputacije[39]

Postupak je sličan kao kod algoritma M1 s tim da se za jedan scenario mora izvršiti imputacija na osnovu više paterna.

Simulacija - priprema skupova [46]

Pored opisanih skupova koje je potrebno kreirati za svaku imputaciju, može se proslediti i skup vrednosti koje su uklonjenje kako bi se izmerila i tačost imputacije. Merenje tačnosti imputacije nije od velikog značaja za ciljeve algoritma ali, ali uzimajući u obzir već navedene zaključke da tačnost imputacije korespondira tačnijem finalnom modelu, radi dodatnih informacija i detaljnije analize se može implementirati na ovom mestu.

5.4.3 Opšte napomene

Struktura koda

Za petlju gde se vrše particije važe slične napomene kao i za prethodni algoritam M1. Ukoliko premeštanje petlje vodi usložnjavanju postupka, a ne dodnosi bitne nove informacije ostaviti postupak što jednostavnijim. Ako se radi o istoj složenosti onda, zavisno od ciljeva i načina praćenja rezultata, se nezavisnim petljama može promenit redosled.

Hiperparametri

Za razliku od prethodnog algoritma M1, primena ovog algoritma se najčešće odnosi na kompletan polazni skup i veliki broj obučavanja modela. Zato se ovde preporučuje izdvajanje pretrage prostora hiperparametara pre početka simulacije. U tom slučaju preporuka je da se prilikom kros-validacije veličina skupova koji s izdvajaju za testiranje odgovara srednjoj vrednosti količina MD koje se planiraju testirati.

Način izvršavanja algoritama

Ukoliko se razmatra veći skup finalnih modela (F) i modela imputacije (Im) izvršavanje algoritma se može razdvojiti na dve faze. U prvoj fazi se mogu izvršiti preliminarni eksperimenti za svaki finalni model i svaki model imputacije za jednostavne parametre simulacije kao što je broj ponavljanja ili skup hiperparametara. Praćenjem dobijenih rezultata se može izvršiti redukcija polaznog skupa modela. Nakon izbora kombinacije modela, elemenata skupa $FxIm$, koji pokazuju najbolje performanse se dalje, u drugooj fazi, može uraditi kompletna simulacija.

5.5 Primena algoritama M1 i M2

U prethodna dva poglavlja je opisana metodologija za pristup analizi uticaja nedostajućih podataka prilikom razvoja prediktivnih modela mašinskog učenja. Metodologija je razdvojena na dva algoritma u skladu sa ciljevima istraživanja u okviru kojih se koristi. Uz svaki algoritam su data posebna i opšta uputsva za njihovu realizaciju koja su značajna za konkretnu realizaciju i koja mogu usmeriti implementaciju u okviru eksperimentalnih istraživanja. Algoritmi, uputstva i napomene su proizišle na osnovu pregleda metodoligija prethodnih istraživanja i na osnovu iskustva u realizaciji eksperimenata u okviru istraživanja ove disertacije, ali su date u opštem obliku. Iako su razvijani i navedeni sa namerom da se metode mašinskog učenjakoriste kao prediktivne metode i metode imputacije mogu se koristiti i u prilikom istraživanja drugih metoda.

Prilikom publikacije novog prediktivnog modela, neophodno je objaviti detalje o nedostajućim podacima ukoliko oni postoje ili ako se smatra da mogu postojati u ulaznom skupu podataka nad kojim će formirati taj prediktivni model. Primena neke od navedenih metodologija omogućava dobijanje informacija uticaju MD, pre svega u odnosu na korišćenje podskupa podataka koje sadrži samo kompletne podatke. U slučaju primene M1, kada se algoritam koristi za razvoj

finalnog modela sa imputacijom, prilikom objavljivanja rezultata neophodno je navesti: opis struktura odnosno paterna MD, metod koji je korišćen za imputaciju i sve zaključke dobijene analizom koji mogu biti od koristi drugom istraživaču. U slučaju primene M2, gde se algoritam koristi u slučaju kad već postoji formirani finalni model, a pretpostavka je da će se on primenjivati i ponovo realizovati nad podacima koji imaju MD, prilikom objavljivanja rezultata treba navesti: sve razmatrane scenarije MD, metode imputacije i rezultate dobijene analizom sa uputstvima i savetima da li pristupati izabranoj imputaciji.

U narednom poglavlju je, na primeru utvrđenog prediktivnog modela, iskorišćen algoritam M2 za analizu uticaja nedostajućih podataka na performanse modela i ponašanja veštačke neuralne mreže kao metode imputacije.

6. Studija slučaja

U ovom poglavlju je, kroz demonstraciju korišćenja opisane metodologije u prethodnom poglavlju, prikazana analiza uticaja imputacija nedostajućih podataka na razvoj prediktivnog modela iz oblasti medicinske dijagnostike. Primenjen je algoritam za simulaciju nedostajućih podataka kod kompletnog skupa prikazan u 5.2 i tom prilikom je razmatrana veštačka neuralna mreža kao metod imputacije.

U nastavku je prvo predstavljen prediktivni model od interesa za koji će se vršiti analiza uticaja MD, a zatim i motivacija da se izvrši ova analiza. Nakon toga je dat opis podataka i metoda koje su korišćene. Pre samih rezulatata aanlize predstavljene su akcije koje su prethodile eksperimentima. Posebno je izdvojeno potpoglavlje koje se odnosi na notaciju koja je korišćena za prikaz eksperimentalnih rezulata. Na kraju su dati rezultati eksperimenata uz odgovarajuće diskusije i zaključke.

6.1 Kontekst problema: izabrani prediktivni model

Sa naglim porastom dostupnih podataka i razvojem računarstva porasla je popularnost istraživanja i primene mašinskog učenja u različitim oblastima pa i u medicini. Metode mašinskog učenja, i nadgledane i nenadgledane, su našle primenu u raznim dijagnostičkim sistemima kao sredstva predikcije određenih stanja i oboljenja [58, 59, 60, 61, 62, 63].

Istraživanje koje smo objavili u [4] predstavlja analizu algoritamske dijagnostike metaboličkog sindorma [64] gde je utvrđeno da se slučajne šume bolje pona-

šaju kao prediktvni model ovog sindroma u odnosu na neuralne mreže korišćenjem istih ulaznih parametara.

Inicijalna motivacija za razvoj ovakvih prediktvnih modela je ta što je predikcija nekog rizika često važnija od trenutne procene rizika [65, 66]. Međutim, faktori rizika često uključuju laboratorijske vrednosti koje nisu uvek dostupne i iziskuju veće troškove. Stoga, u medicinskoj dijagnostici, postoji interes da se razviju prediktivni modeli koji koriste samo jednostavne, osnovne i lako dostupne podatke.

U navedenoj studiji smo utvrdili da RF postiže bolje performanse, specifičnost, senzitivnost, preciznost i negativnu prediktvnu vrednost, u odnosu na predikcije realizovane preko LR, DT i ANN. Svi detalji izvršenih eksperimenata za predikciju metaboličkog sindroma preko ANN i RF, kao i diskusija, zaključci i ograničenja su navedeni u datoj publikaciji [4].

6.2 Analiza problema: nedostajući podaci

Što se tiče nedostajućih podataka (MD), medicinski podaci ne predstavljaju izuzetak i često sadrže MD koji imaju različitu strukturu, tip i šablon ili mehanizam na osnovu kog se desilo njihovo nastajanje [3, 67].

Prethodna istraživanja, za procenu metaboličkog sindroma, su vršena nad kompletnim podacima. Korišćenjem dobijenih saznanja iz analize imputacije preko ANN prikazane u 4.4 i novog polaznog skupa podataka, u nastavku istraživanja je izvršena analizu uticaja potencijalnih MD na rezultate upotrebe slučajnih šuma u predikciji metaboličkog sindroma.

Izvršena je analiza uticaja imputacija preko ANN za:

1. MD kod laboratorijskih vrednosti koje se koriste samo u fazi pretprocesiranja,
2. MD kod prediktora koji se koriste u finalnom prediktivnom modelu.

Oba slučaja su razmatrana sa ciljem da se utvrdi da li neka nedostajanja i dalje mogu obezbediti izgradnju finalnih prediktivnih modela upoređujujući rezultate sa slučajem kad bi se za razvoj koristio samo podskup sa kompletnim podacima (CC).

Posebno je naglašena razlika između ova dva slučaja s obzirom da se scenariji MD, kao što je (1), retko analiziraju u literaturi, a najveći fokus u istraživanjima

	Mean	St. Dev.	Min	Max
GEN	1.463	0.499	1	2
AGE	43.413	10.615	18	69
BMI	29.732	6.472	16.600	50.440
WC	96.499	14.814	59.600	153.400
WHtR	0.565	0.091	0.338	0.899
SBP	132.984	18.287	92	210
DBP	85.936	12.839	55	137
HDL	1.124	0.262	0.460	2.090
TG	2.057	1.819	0.350	27.320
GLY	5.145	1.321	2.800	13.800

Tabela 6.1: Deskriptivna statistika podataka koji se koriste u istraživanju

je usmeren na analizu uticaja MD kod prediktora (2). Takođe, prilikom primene imputacija se izlazne vrednosti često zanemariju kao vrednosti koje mogu učestvovati u procesu imputacije što je u ovom istraživanju upravo urađeno kroz analizu slučaja (1).

6.3 Okvir istraživanja: podaci, metode i postavke

Podaci koji su korišćeni predstavljaju uzorke nastale na Klinici za endokrinologiju, dijabetes i bolesti metabolizma Kliničkog centra Vojvodine. Obuhvataju antropološke i laboratorijske podatke koji se koriste za analizu kardiometaboličkog rizika i metaboličkog sindroma. Svi podaci su prikupljeni u skladu sa Helsinškom deklaracijom. Rešenjem 00-1024 Etičkog odbora kliničkog centra je odbrena upotreba podataka u svrhe ovog istraživanja. Odluka o saglasnosti se nalazi u prilogu A.6.

Baza podataka sadrži sledeće podatke: pol (GEN), uzrast (AGE), indeks telesne mase (BMI), obim struka (WC), odnos obima struka i visine (WHtR), sistolni pritisak (SBP), dijastolni pritisak (DBP), LDL holesterol (LDL), HDL holesterol (HDL), ukupni holesterol (TCH), trigliceridi (TG) i glikemija (GLY). Deskriptivna statistika podataka je prikazana u tabeli 6.1.

Prilikom direktne procene metaboličkog rizika koriste se laboratorijski podaci i vrednost procene (MetS) se dobija na osnovu izračunavanja prikazanog na

listingu A.1 u dodatku što podrazumeva postojanje vrednosti za HDL, TG i GLY. Za algoritamsku procenu rizika korišćenjem algoritma RF se podrazumevaja postojanje vrednosti za GEN, AGE, BMI, WHtR, SBP i DBP.

Svi eksperimenti su izvršeni korišćenjem programskog jezika R , u okruženju RStudio uz korišćenje različitih biblioteka. Za realizaciju ANN je korišćena implementacija *rprop+* [47] algoritma iz biblioteke *neuralnet* [68], a za RF implementacija iz biblioteke *randomForest* [69].

Postavke za ANN koja se koristila kao metod imputacije su:

- algoritam: rprop+ (resilient backpropagation) [47, 48],
- aktivaciona funkcija: log i tanh (logistic i tangent hyperbolicus),
- inicijalne težine: slučajna inicijalizacija,
- stepmax: različito za svaki eksperiment,
- threshold (vrednost parcijalnog izvoda funkcije greške): različito za svaki eksperiment,
- kriterijum zaustavljanja: bilo koji od parametara zaustavljanja (stepmax i threshold) koji se pre dostigne,
- tip predikcije: regresija (linear output).

Ulazni i izlazni skupovi prilikom obučavanja ovih modela su se razlikovali u zavisnosti od scenarija i paterna MD koji se u njima pojavljuju. Ulazne vrednosti su predstavljale podskup vrednosti promenljivih koje imaju dostupne podatke, a izlazni skupovi su predstavljali vrednosti promenljivih kod kojih nedostaju podaci.

Postavke za RF koja se koristila za finalni prediktivni model su:

- algoritam: Breiman [70],
- broj slučajno izbaranih promenljivih za stabla: 1/3 ukupnog broja promenljivih,
- broj stabala: 100,
- veličina uzorka za stabla: 63,2%,
- tip predikcije: klasifikacija.

Ulazni i izlazni skupovi za ovaj model odgovaraju po strukturi skupovima koji su korišćeni za razvoj modela prikazan u publikaciji [4], odnosno ulazni skupovi se odnose na vrednosti nelaboratorijskih podataka, a izlazna vrednost predstavlja procenu postojanja metaboličkog sindroma (MetS).

Prilikom podele skupova su se koristile funkcije paketa *caret* [49] koje, između ostalog, omogućavaju i podelu skupova tako da se održi balansiranost podataka u

odnosu na neku klasu što je i iskorišćeno prilikom realizacije simulacija u odnosu na MetS (stratifikacija).

6.4 Realizacija istraživanja: eksperimentalni rezultati

Korišćenjem algoritma M2 izvršena je analiza uticaja imputacija za 21 slučaj MD. Prikaz razmatranih slučajeva je dat u tabeli 6.2. U tabeli je, za svaki posmatrani slučaj, prikazano: kod kojih vrednosti se pojavljuju MD, kom tipu eksperimenta pripada, koliko paterna MD podrazumeva posmatrani slučaj, po kom mehanizmu nastaju MD i koja je učestalost svakog postojećeg paterna MD u okviru jednog scenarija. Karakterom "/" je označeno da se MD javljaju nezavisno jedni od drugih, dok je karakterom "+" označeno istovremeno nedostajanje.

Postupak ispitivanja je razdvojen prema broju promenljivih kod kojih se javljaju MD što određuje tip eksperimenta:

1. kada podaci nedostaju samo za jednu promenljivu (UVmd),
2. kada podaci nedostaju za više promenljivih (MVmd).

Slučajevi iz grupe (1) se obično ređe javljaju u realnim situacijama i uglavnom se javljaju na manjim skupovima podataka ili kod podataka gde postoji manji broj promenljivih. Međutim, analiza ovakvih slučajeva može dati preciznije uvide u ponašanje imputacija i njihov međusobni uticaj. Iz tog razloga je za takve slučajeve urađena analiza koja je obuhvatila scenarije za sva tri mehanizma nedostajanja (MCAR, MAR, MNAR) i različite količine MD. Za eksperimente iz grupe (2) je izabrano nekoliko smislenih scenarija za koje se može pretpostaviti pojavljivanje i koji mogu ukazati na potencijalnu pojavu problema pri postojanju MD.

Parametri algoritma M2 koji su se koristili prilikom analize su dati u nastavku.

Skup finalnih prediktivnih modela F sadrži jedan elemenat i to je RF model za predikciju metaboličkog rizika.

Skup razmatranih scenarija S sadrži 105 elemenata. Razmatran je 21 slučaj pojavljivanja MD gde su za svaki slučaj analizirani scenariji za količine MD od 10%, 20%, 30%, 50% i 80%.

	Slučaj MD	Tip	Broj paterna	Mehanizam	Učestalost
1	HDL	UV	1	MCAR	1
2	HDL	UV	1	MAR	1
3	HDL	UV	1	MNAR	1
4	TG	UV	1	MCAR	1
5	TG	UV	1	MAR	1
6	TG	UV	1	MNAR	1
7	GLY	UV	1	MCAR	1
8	GLY	UV	1	MAR	1
9	GLY	UV	1	MNAR	1
10	HDL/TG/GLY	MV	3	MCAR	(0.3, 0.3, 0.3)
11	HDL/TG/GLY	MV	3	MAR	(0.3, 0.3, 0.3)
12	HDL/TG/GLY	MV	3	MNAR	(0.3, 0.3, 0.3)
13	WHtR	UV	1	MCAR	1
14	WHtR	UV	1	MAR	1
15	WHtR	UV	1	MNAR	1
16	BMI	UV	1	MCAR	1
17	BMI	UV	1	MAR	1
18	BMI	UV	1	MNAR	1
19	WHTR/ SBP+DBP	MV	2	MAR	(0.6, 0.4)
20	SBP+DBP/ WHtR/ WHtR,SBP+DBP	MV	3	MAR,MNAR	(0.4, 0.4, 0.2)
21	BMI+WHtR/ SBP+DBP/ BMI+WHtR+SBP+DBP	MV	3	MAR,MNAR	(0.4, 0.4, 0.2)

Tabela 6.2: Razmatrani slučajevi MD za koje je izvršena analiza uticaja imputacije

Skup svih paterna P sadrži sve paterne nedostajanja koji se pojavljuju u svim scenarijama. Ukupno je analizirano 11 različitih paterna.

Skup metoda imputacije Im sadrži jedan elemenat i to je ANN *rprop* algoritam i odnosi se na klasu ovih modela. Konkretna instanca ANN modela koja je obučena u simulaciji i koja se koristila prilikom same imputacije zavisi od ulaznog i izlaznog skupa i hiperparametara koji odgovaraju određenim paternima nedostajanja o čemu će biti reči u nastavku. U skladu sa rezultatima istraživanja opisanim u 4.4 ANN je izabrana kao metod imputacije s obzirom na sličnost domena i skupova podataka koji su se koristili u istraživanjima.

Skup parametara simulacije SP čine sledeći elementi:

- broj simulacija istovetnih amputacija: 20,
- broj slojeva kCV: 10,
- broj ponavljanja kCV: 5.

Ovo znači da je za svaki scenario, od 105 ukupno, izvršeno 1000 simulacija imputacija MD koje podrazumevaju obučavanje jedne ili više odgovarajućih ANN i isto toliko obučavanja finalnog prediktivnog modela nad dobijenim, popunjenim skupovima.

6.4.1 Priprema modela za imputaciju

Pre same simulacije određeni su hiperparametri za obučavanje ANN.

Prvo su razmatrane strukture ANN za imputacije laboratorijskih vrednosti, odnosno za vrednosti promenljivih koje se koriste samo za izračunavanje izlaznih vrednosti pre formiranja finalnog prediktivnog modela. Unakrsnom validacijom je izvršeno pretraživanje hiperparametara za pronalaženje optimalnog skupa parametara koji će kasnije biti korišćeni prilikom imputacije vrednosti za HDL, TG i GLY.

Grid pretraga je podrazumevala sledeće: broj neurona u skrivenom sloju je postavljan na: 2, 4, 6, 8 i 10; parametar *threshold* na: 0.01 i 0.001; parametar *stepmax* je fiksiran na 5000. Ove vrednosti su određene prethodnim, inicijalnim testiranjima. Nije vršeno merenje performansi na spoljašnjem izdvojenom test skupu podataka jer se ovom prilikom nisu formirali niti testirali modeli koji bi se koristili kao finalni prediktivni modeli. Validacija je realizovana kao kCV sa ponavljanjem gde je broj slojeva postavljen na 5, a broj ponavljanja na 10. S obzirom da je polazni

skup deljen na 5 disjunktnih skupova za obučavanje što bi odgovaralo slučaju kada u polaznom skupu postoji 20% nedostajućih podataka.

Zavisno od paterna MD svaka od promenljivih (HDL, TG, GLY) se može pojaviti samostalno kao promenljiva sa MD ili zajedno sa nekom drugom promenljivom. Slučaj kada su sve tri laboratorijske vrednosti nedostajale nije razmatran jer se time praktično problem svodi na polazni, odnosno na predikciju MetS na osnovu nelaboratorijskih vrednosti i može se posmatrati kao zaseban test slučaj finalnog modela.

Dalje, model za imputaciju jedne ovakve vrednosti može biti realizovan na dva načina:

- koristeći samo prediktore za ulazni skup,
- koristeći sve dostupne informacije, a to podrazumeva i ostale dostupne laboratorijske vrednosti, kao ulazni skup.

Implementacija rešenja zasnovanog na prvom predlogu bi bila znatno jednostavnija, a izvršavanje manje složeno. Međutim, time bi se problem sveo na pitanje: da li je suma delova veća od celine, odnosno da li se predikcija MetS-a može uraditi kao rezultat predikcije njenih laboratorijskih vrednosti. Time ovo spada u domen odabira finalnog modela za MetS što nije predmet ovog rada. Iz tog razloga koristimo drugi metod, koji jeste složeniji, ali odgovara cilju istraživanja: utvrditi kako se poznati model menja ukoliko postoje MD. Zbog toga je pretraga parametara za svaku promenljivu podrazumevala pretragu za tri slučaja korišćenja zavisno od toga da li ostale promenljive imaju dostupne vrednosti ili ne. Svi rezultati simulacije su prikazni na slikama A.1, A.3 i A.2 i tabeli A.1 koje su date u dodatku.

Ono što se može primetiti je da ni u jednom slučaju ne postoje značajne razlike u izboru parametara pogotovo u odnosu na vrednost parametra *threshold*. Iz tog razloga je za ovaj parametar izabrana vrednost 0.01 za buduće imputacije. Što se tiče broja neurona rezultati zavise od međusobnog odnosa među promenljivama. Za slučaj HDL promenljive se može videti da TG utiče na procenu HDL vrednosti što je i očekivano. Na osnovu trenda pada u rezultatima može se pretpostaviti da bi povećanje broja neurona dovelo do bolje procene, pa je za slučajeve gde učestvuje TG uzeto da je broj neurona 10, a za slučaj kad nedostaje TG broj skrivenih neurona je 2. Za TG vrednosti se primećuje da je najlošija procena kad ne učestvuje nijedna dodatna laboratorijska mera, a da HDL ima tendenciju da popravi procenu TG što je u skladu i sa prethodnim razmatranjem. Zato, za slučajeve kad učestvuje HDL, je uzeto da je broj neurona 10, a u suprotnom 4. Sa poslednje slike i u odnosu na

prethodne zaključke se vidi da HDL ima manji uticaj nego TG na GLY pa je za slučajeve kad TG učestvuje uzeto da je broj neurona 10, a u suprotnom 8.

Konačni izbor parametara koji su učestvovali u imputacijama je prikazan u tabeli 6.3.

Promenljiva	Ulaz za ANN	HN
HDL	p + GLY, TG	10
HDL	p + GLY	2
HDL	p + TG	10
TG	p + GLY, HDL	10
TG	p + GLY	4
TG	p + HDL	10
GLY	p + TG, HDL	10
GLY	p + TG	10
GLY	p + HDL	8
threshold = 0.01, stepmax = 5000		

Tabela 6.3: Konačni izbor parametara za ANN namenjenih za korišćenje prilikom simulacija imputacija nedostajućih podataka

Drugi deo analize imputacija se odnosi na imputacije vrednosti koji se odnose na prediktore (nelaboratorijski podaci). S ozbirom na potencijalne različite paterne koji se mogu javiti i preklapanja nedostajanja kod nekih promenljivih za različite paterne nije pretražen ceo skup hiperparametara koji odgovara podskupovima ulaznih vrednosti već je izvršena pretraga za model koji odgovara proceni WHtR vrednosti, kao značajne promenljive pa ja za sve ostale prediktore fiksiran broj skrivenih neurona na 10. Kao što je navedeno u uputstvima za algoritam, precizna procena hiperparametara nije neohodna za ovu vrstu analize jer cilj istraživanja nije formiranje finalnog konačnog modela. Tek ukoliko se nakon analize, izabere metod i utvrdi opravdanost imputacije za konkretan skup podataka tada je preporučeno uraditi i preciznu procenu hiperparametara.

6.4.2 Notacija za prikaz rezultata

Za potrebe amputacije, a radi jednostavnijeg opisa scenarija u nastavku su uvedene notacije koje odgovaraju i realizovanoj implementaciji. Notacije se, pre svega, odnose na parametre koji se koriste u procesima amputacije kroz korišćenje funkcije *ampute* [52, 53] i odgovaraju strukturama opisanim u 5.2.1.

Za definisanje paterna se pretpostavlja da se uvek polazi od paterna gde nijedna promenljiva ne nedostaje odnosno od jediničnog vektora u kom svaki član predstavlja jednu promenljivu. Označićemo ovaj jedinični vektor sa *pattern.boot*.

Specifičan patern se definiše tako što se određene komponente postavljaju na 0 čime se ukazuje da vrednosti za tu promenljivu nedostaju. Na primer, vektor

$$(1, 1, 0, 1, 1, 1, 1, 0, 1, 0)$$

bi označavao da nedostaju vrednosti za BMI, HDL i GLY.

Radi lakše čitljivost i sledeći prepoznatljive programske oznake za vektore, u tekstu će se ovaj patern navoditi kao

$$p(BMI, HDL, GLY) = 0.$$

Zapis opšteg oblika paterna će, stoga, biti:

$$p(var_1, var_2, ..., var_{pn}) = 0$$

gde je var_i promenljiva kod koje postoji nedostajanje, a pn - broj promenljivih kod kojih postoji nedostajanje.

Ukoliko scenario sadrži više paterna neophodno je definisati matrice čije vrste predstavljaju paterne. Na primer, matrica

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

označava da scenario podrazumeva dva paterna: jedan gde nedostaju BMI, HDL i GLY, a drugi gde nedostaju SBP i DBP.

U tekstu će se matrica koja predstavlja navedene paterne pisati kao:

$$mp[p(BMI, HDL, GLY) = 0, p(SBP, DBP) = 0]$$

gde mp označava da se radi o matrici sa vektorima vrsta definisanim navedenim paternima.

Tako će zapis opšteg oblika matrice sa paternima biti:

$$misPattern : mp[p_1, p_2, ..., p_{mpn}]$$

gde p_i predstavlja jedan patern, a mpn broj paterna u razmatranom scenariju.

U slučaju MAR i MNAR mehanizma za svaki patern je neophodno navesti i odgovarajuće težine na osnovu kojih se donosi odluka o amputaciji. Uvek se polazi od nula-vektora odnosno pretpostavlja mse da nijedna promenljiva ne utiče na MD. Označićemo ovaj polazni vektor težina sa *weights.boot*.

Postavljanjem odgovarajućih vrednosti za određene komponente kreiraju se različiti mehanizmi MD za vrednosti koje su definisan paternima MD. Na primer, vektorom

$$(0.5, 1, 1, 0, 0, 0, 0, 0, 0, 0)$$

bi se modelirali MAR ili MNAR mehanizam gde GEN, AGE i BMI utiču na nedostajanje i to GEN upola manje nego AGE i BMI.

U tekstu će ovi vektori težina biti navedeni kao

$$w(GEN, AGE, BMI) = (0.5, 1, 1).$$

Zapis opšteg vektora sa težinama koji se odnosi na jedan patern će biti:

$$w(var_1, var_2, ..., var_{wn}) = (value_1, value_2, ..., value_{wn})$$

gde var_i označava promenljivu čije vrednosti utiču na nedostajanje drugih vrednosti, $value_i$ predstavlja težinski faktor tog uticaja, a wn je broj promenljivih koje utiču na realizaciju mehanizma nedostajanja.

Za svaki patern u jednom scenariju je potrebno definisati odgovarajući vektor težina kao što je iznad definisano. Ti vektori predstavljaju vrste matrice težina. Na primer, za već spomenuti scenario, matrica težina bi mogla biti:

$$\begin{pmatrix} 0.5 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

U tekstu će ove matrice težina biti zapisane kao:

$$mw[w(GEN, AGE, BMI) = (0.5, 1, 1), w(SBP, DBP) = (1, 1)]$$

gde mw označava da se radi o matrici čiji su vektori vrsta definisani zadatim vektorima težina.

Tako će zapis opšteg oblik matrice težina biti:

$$misWeights : mw[w_1, w_2, ..., w_{mpn}]$$

gde je w_i vektor težina, a mpn broj paterna u razmatranom scenariju.

U navedenom primeru, prvi vektor težina je zadat tako da odgovara MAR mehanizmu, a drugi vektor odgovara MNAR mehanizmu. Osim modeliranja MAR i MNAR mehanizama, moguće je modelirati i MCAR mehanizam ukoliko se sve težine u vektoru postave na 0 (*weights.boot*). Ovde se može primetiti na koji način predložene strukture daju mogućnost krieranja najrazličitijih struktura MD u okviru jednog scenarija.

Oznakom $mech$ se označava osnovni mehanizam nedostajanja nekog scenarija ukoliko nisu definisani težinski vektori za izračunavanje težinskog skora na osnovu kog se vrši izbor podataka za amputaciju. Dozvoljene vrednosti za $mech$ su: MCAR, MAR i MNAR, a podrazumevana vrednost je MAR.

Osim težina, zadaje se i raspodela po kojoj će se na osnovu težina birati podaci za uklanjanje. Na primer, vektor

$$(3, 2, 1, 0)$$

označava da se će se na osnovu težinske sume skup podataka podeliti u četiri grupe. Vrednosti navedenog vektora označavaju relativne verovatnoće (eng. odds) na osnovu kojih se određuje da li će vrednost biti uklonjena ili ne. Dati primer označava da će najviše biti nedostajanja za najmanju vrednost težinske sume dok najveće vrednosti neće biti uklonjene.

Opšti oblik raspodele MD po grupama za jedan patern se zadaje u obliku:

$$o(value_1, value_2, ..., value_{on})$$

gde je $value_i$ relativna verovatnoća nedostajanja u za odgovarajuću grupu, a on broj grupa na koje se dele podaci na osnovu vrednosti težinske sume pri čemu veće vrednosti sume odgovaraju većim vrednostima za i.

Treba voditi računa da broj grupa bude optimalan u odnosu na ostale zadate vrednosti jer je nekad nemoguće izvršiti podelu skupa u skladu sa zadatim relativnim verovatnoćama. Svakako, algoritam amputacije neće proizvesti grešku već će u tom slučaju uzeti najveći mogući broj grupa i po tome izvršiti amputaciju.

Ukoliko scenario ima više paterna, opštem obliku odgovara matrica:

$$misOdds\ :\ mo[o_1, o_2, ..., o_{mon}]$$

gde je o_i raspodela pojavljivanja MD na osnovu težinskih vrednosti za patern p_i.

Na kraju, ukoliko scenario ima više paterna potrebno je definisati učestalost pojave svakog od njih. Na primer, vektor

$$(0.4, 0.6)$$

ukazuje na to da se prvi patern pojavljuje u manje slučajeva (0.4) nego drugi (0.6). Zbir svih vrednosti ovog vektora mora biti 1.

Opšti oblik učestalosti je sledeći vektor:

$$misFreq\ :\ mf(val_1, val_2, ..., val_{mfn})$$

gde je val_i učestalost paterna p_i, mpn predstavlja broj paterna u posmatranom scenariju i važi da je $\sum_{n=1}^{mfn} val_i = 1$.

6.4.3 Prikaz rezultata

U nastavku će, prvo, na primeru jedne promenljive biti detaljno opisan način na koji su se formirali i analizirali posmatrani scenariji MD. Nakon toga su prikazni rezultati i diskusija za svaki scenario pojedinačno.

Scenario: UVmd - HDL

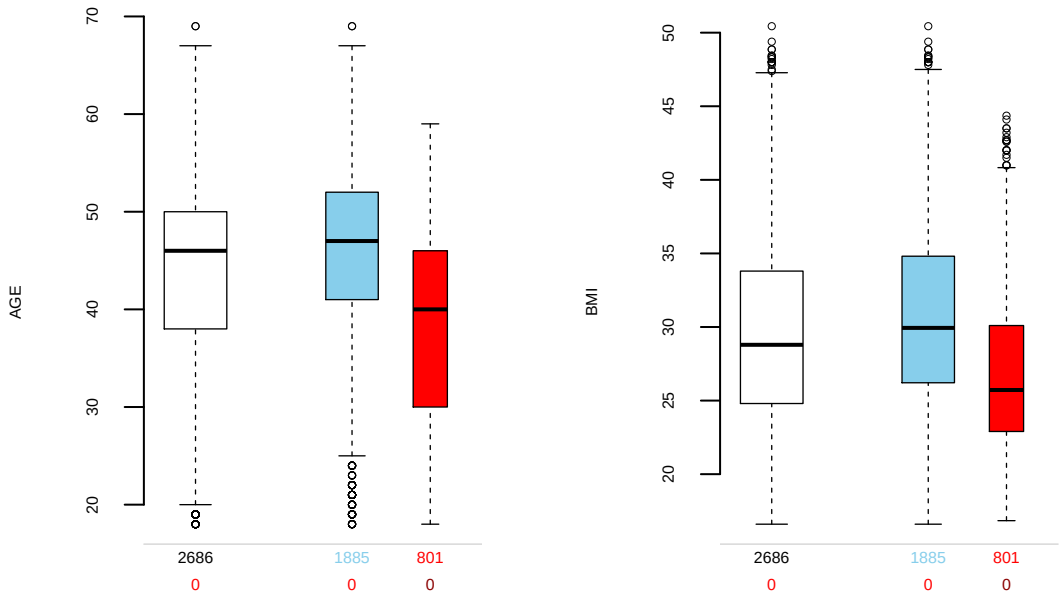
Posmatrajmo prvo scenarije kada nedostaju vrednosti za jednu promenljivu, HDL, po MAR mehanizmu. Scenariji koji su izvršeni za ovaj slučaj su:

- UVmd HDL MAR 0.1,
- UVmd HDL MAR 0.2,
- UVmd HDL MAR 0.3,
- UVmd HDL MAR 0.5,
- UVmd HDL MAR 0.8.

Postavke koje važe za svaki UVmd HDL MAR scenario su sledeće:

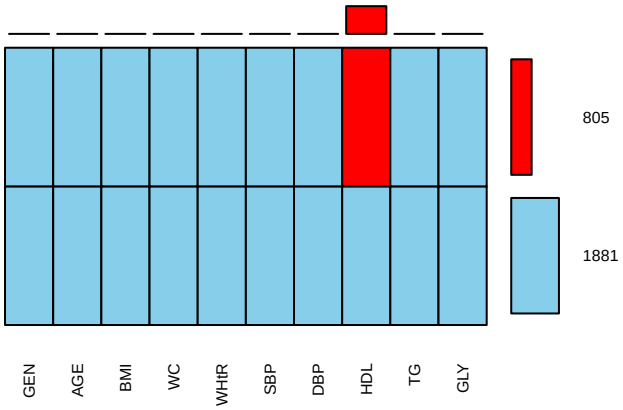
- $mech\ :\ MAR$,
- $misPattern\ :\ mp[p(HDL) = 0]$,
- $misWeights\ :\ mw[w(AGE, BMI) = (1, 0.8)]$,
- $misOdds\ :\ mo[o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1)]$ za količine od 50% i 80%,
- $misFreq\ :\ mf(1)$.

Postavljanjem težina i pravila za distribuciju definisan je, u stvari, sledeći mehanizam: vrednosti za HDL promenljivu nedostaju zavisno od godina i indeksa telesne mase, s tim da godine imaju malo veći uticaj na pojavu nedostajanja. Ovo je postignuto postavljanjem vrednosti za parametar *missWights*. Parametrom *missOdds* je definisano da će biti više podataka koji nedostaju za manje vrednosti ovih parametara. Ovako definisanim paternom je, praktično, modelovana ideja da mlađi i manje gojazni pacijenti ređe vrše laboratorijske kontrole holesterola. Ovo se može videti i na slici 6.1 gde je prikazana razlika u distribuciji vrednosti AGE i BMI za kompletne podatke i one za koje nedostaju vrednost HDL.



Slika 6.1: Distribucija vrednosti za AGE i BMI u početnom skupu i nakon amputacije za scenario gde 30% HDL vrednosti nedostaje po MAR mehanizmu zavisno od AGE i BMI

Na slici 6.2 se vidi primer ovog jednostavnog paterna sa brojem opservacija za količinu od 30% nedostajanja na jednom primeru uklanjanja podataka (amputacije). U primeru je izvršena imputacija za 2686 podataka što odgovara 90% polaznog skupa s obzirom da će se u simulaciji izdvajati 10% podataka za finalno testiranje prediktivnog modela.



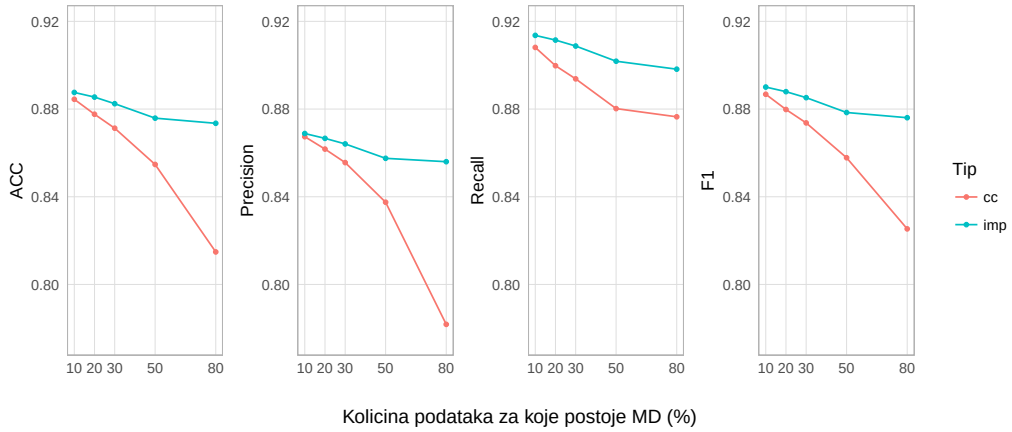
Slika 6.2: HDL MAR 0.3 Pattern

Rezultati svih simulacija su praćeni posmatranjem performansi finalnih prediktivnih modela i to na osnovu sledećih mera: Accuracy, Kappa, Precision, Recall, Specificity i F1. Iako su vrednosti za meru Kappa evidentirane u nastavku nije posebno diskutovana jer njene promene prate promene ostalih mera perfomansi.

Prosečne vrednosti ostvarene tokom svih scenarija za opisani slučaj (UVmd HDL MAR) su prikazani u tabeli 6.4 i na slici 6.3.

MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
10	cc	0.884	0.768	0.867	0.908	0.862	0.887
10	imp	0.888	0.775	0.869	0.914	0.864	0.890
20	cc	0.878	0.755	0.862	0.900	0.857	0.880
20	imp	0.885	0.770	0.867	0.911	0.861	0.888
30	cc	0.871	0.742	0.856	0.894	0.851	0.874
30	imp	0.882	0.764	0.864	0.909	0.858	0.885
50	cc	0.855	0.708	0.838	0.880	0.830	0.858
50	imp	0.876	0.751	0.858	0.902	0.852	0.878
80	cc	0.815	0.628	0.782	0.876	0.754	0.825
80	imp	0.874	0.746	0.856	0.898	0.850	0.876

Tabela 6.4: Scenario: **UVmd HDL MAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije



Slika 6.3: Scenario: **UVmd HDL MAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije

Iz rezultata se može ustanoviti da imputacija podataka poboljšava performanse finalnih modela u odnosu na formiranje modela nad kompletnim podacima. Može se primetiti da je odziv veći nego preciznost . Nivelacija te dve vrednosti se ogleda kroz meru F1. Ono što se može primetiti je da imputacija značajnije poboljšava preciznost nego odziv za veće količine MD pa se i ukupna performansa modela sa imputacijom znatno povećava za veće količine MD (ACC i F1). Sličnost rezultata za ACC i F1 se može pripisati balansiranosti polaznog skupa i očuvanju te balansiranosti kroz podelu skupa za obučavanje uz korišćenje stratifikacije.

Na sličan način su izvršene simulacije za MCAR i MNAR mehanizam sa sledećim razlikama.

Za MCAR mehanizam se ne definišu vrednosti za težinske parametre (*missWeights*) jer je nedostajanje slučajno i ne zavisi ni od jedne druge veličine. Iz istog razloga parametar *misOdds* nema vrednost. Tako su postavke za MCAR mehanizam sledeće:

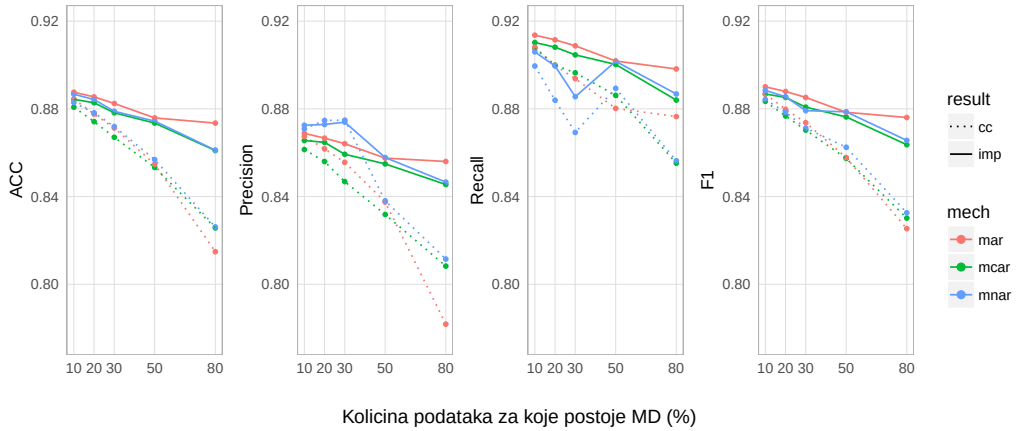
- $mech : MCAR$,
- $misPattern : mp[p(HDL) = 0)]$,
- $misWeights : weights.boot$,
- $misOdds : null$,
- $freq : mf(1)$.

Za MNAR mehanizam se težinski vektor menja tako da za komponentu koja odgovara promenljivoj koja se posmatra ima vrednost različitu od 0. To, na primer, za promenljivu HDL znači postavljanje sledećih postavki:

- $mech$: $MNAR$,
- $misPattern$: $mp[p[HDL] = 0]$,
- $misWeights$: $mw[w(HDL) = 1]$,
- $misOdds$: $mo[o(0, 1, 2, 3)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1)]$ za količine od 50% i 80%,
- $misFreq$: $mf(1)$.

Raspodela MD vrednosti je postavljena tako da kod većih vrednosti postoji više nedostajanja jer opet pokušavamo da modeliramo situaciju u kojoj potencijalno zdraviji pacijenti ređe vrše kontrole. S obzirom da je HDL takozvani "dobar holesterol" on je obrnuto srazmeran oboljenju pa uklanjamo pacijente sa višim vrednostima.

Rezultati za sve realizovane scenarije za promenljivu HDL su navedeni u tabeli A.2 (u dodatku) a prikazani su na slici 6.4.



Slika 6.4: Scenariji: **UVmd HDL - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

Primećujemo da su trendovi u poboljšanju performansi finalnog modela slični za sva tri scenarija. Ovo je dobra indikacija za nastavak istraživanja jer se manifsetuje određena neosetljivost na mehanizam nedostajanja.

Ono što se takođe može primetiti da kod MNAR mehanizma, za slučaj gde 50% podataka ima neku vrednost koja nedostaje, dolazi do određenih poboljšanja odziva u odnosu na slučaj sa 30% MD. Međutim, promene u preciznosti takođe prate ovu pojavu pa se i ta uobičajena razmena uticaja ovih mera može videti kroz usaglašene rezultate za F1 i ACC. S obzirom da i model nad skupom sa kompletnim podacima pokazuje isti trend možemo pretpostaviti da je ovom uzrok raspodela vrednosti i realizacija simulacija i to na potencijalno dva mesta: inicijalna podela

skupova za spoljnu CV (stratifikacija po MEtS) i distribucija amputiranih MD. Iako ova pojava nije od značaja za ciljeve ove analize mogli bi se eventualno ponoviti eksperimenti sa različitim inicijalnim podelama skupova (drugačiji niz slučajnih brojeva sa ili bez stratifikacije) i pojednostavljivanjem distribucije pojave MD za veće količine. Na primer, $misOdds$ parametra bi se mogao postaviti na $c(1)$ umesto $c(2, 1)$.

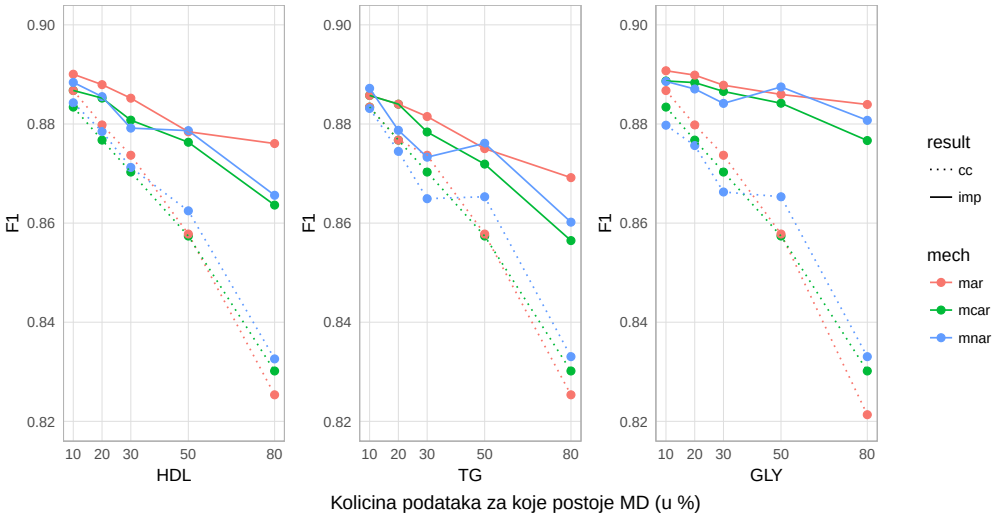
Opšti zaključci, u skladu sa ciljevima simulacije, su:

- Ukoliko novi skup podataka ima MD kod HDL promenljive treba pristupiti imputaciji pre formiranja finalnog prediktivnog modela.
- Ukoliko je polazni skup bio takav da su iz njega uklonjeni podaci koji sadrže MD kod HDL onda je to uklanjanje možda nepotrebno dovelo do lošijeg finalnog prediktivnog modela.

Scenariji: UVmd - HDL, TG, GLY

Prethodno su opisani rezultati za imputaciju kad vrednosti nedostaju za HDL promenljivu. Sličan proces je primenjen i za druge dve spoljašnje laboratorijske vrednosti (TG i GLY).

Svi rezultati su navedeni u tabeli A.2 i prikazani na slikama A.4 i A.5 koje se nalaze u dodatku. Ovde, na slici 6.5 su radi preglednosti, prikazani samo F1 rezultati za sve tri promenljive.



Slika 6.5: Scenariji: **UVmd HDL/TG/GLY - MCAR/MAR/MNAR**. Prikaz F1 rezultata za sve scenarije

Zaključci su slični kao prethodno izneti za slučaj HDL. Imputacija poboljšava finalni model za MD kod sve tri promenljive i nema značajne razlike u odnosu na mehanizam nedostajanja odnosno mehanizam ima uticaja podjednako na formiranje modela kod kompletnog skupa i onog sa imputacijama.

Razlika u performansama raste sa porastom redova koji sadrže MD što je i očekivano. Uviđamo da postoji trend pogoršanja rezultata i kod rezultata nad skupom sa imputacijom, ali je taj trend pogoršanja znatno manji nego kod korišćenja kompletnog skupa. Ovo je posebno uočljivo za GLY gde se dobijaju veoma dobri rezultati čak i za velik broj MD i za svaki mehanizam nedostajanja. To nam ukazuje na to da je ili GLY od manjeg značaja ili se dobro objašnjava ostalim vrednostima i da bi verovatno bilo pogrešno ili makar nepotrebno ukloniti tu promenljivu iz daljeg istraživanja ukoliko bismo primetili da ima puno vrednosti koje nedostaju.

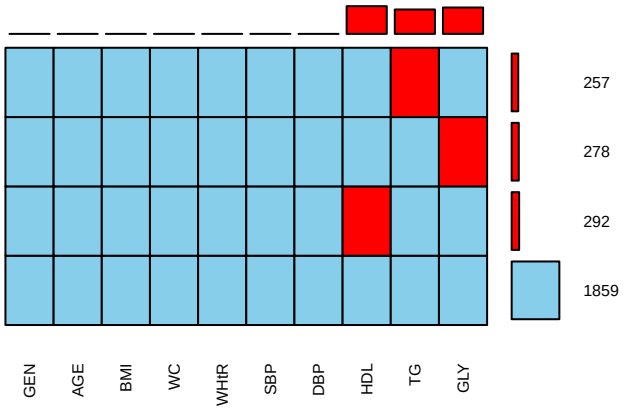
Dalje možemo primetiti i da imputacija kod MAR mehanizma najbolje popravlja predikciju. To znači da ako zaista mlađe osobe sa manjim BMI imaju laboratorijske podatke koji nedostaju nema potrebe uklanjati ih iz daljeg razmatranja.

Takođe, možemo utvriti da se za promenljive TG i GLY dešava slična situacija kod MNAR mehanizma za količine od 30% i 50% MD kao kod HDL vrednosti. Ovo ide u prilog gore navedenoj pretpostavici da je ova pojava uslovljenja parametrima izvršenih simulacija.

Kao opšti zaključak utvrđujemo da finalni modeli koji se razvijaju nad podacima kod kojih je izvršena imputacija daju bolje rezultate nego kad se koriste samo kompletni slučajevi za sve scenarije, da performanse zavise od količine nedostajanja i to drugačije za svaku promenljivu, da je imputacija slabo osetljiva na mehanizam nedostajanja i da za konkretan slučaj MAR mehanizma (manje vrednosti AGE i BMI uslovljavaju pojavu MD) dolazi do najvećeg poboljšanja. Kao što je već navedeno kod HDL slučaja, treba pristupiti imputaciji vrednosti pre formiranja modela, a ukoliko je polazni skup bio takav da su iz njega uklonjeni podaci koji sadrže MD onda je to uklanjanje bilo pogrešno, pogotovo za GLY vrednosti.

Scenariji: MVmd - HDL, TG, GLY

Dalje je razmatran slučaj kad vrednosti mogu nedostajati za bilo koje od tri spoljašnje promenljive koje predstavljaju laboratorijske vrednosti, nezavisno. Na slici 6.6 je prikazan primer sa paternima koji odgovaraju scenariju gde podaci nedostaju po MAR mehanizmu i to kod 30% podataka.



Slika 6.6: Scenariji: **HDL,TG,GLY MAR 0.3 Pattern**

Postavke za scenarije gde MD nastaju po MCAR mehanizmu su:

- $mech$: $MCAR$,
- $misPattern$: $mp[p(HDL) = 0, p(TG) = 0, p(GLY) = 0)]$,
- $misWeights$: $weights.boot$,
- $misOdds$: $null$,
- $freq$: $mf(0.33, 0.33, 0.33)$.

Postavke za scenarije gde MD nastaju po MAR mehanizmu su:

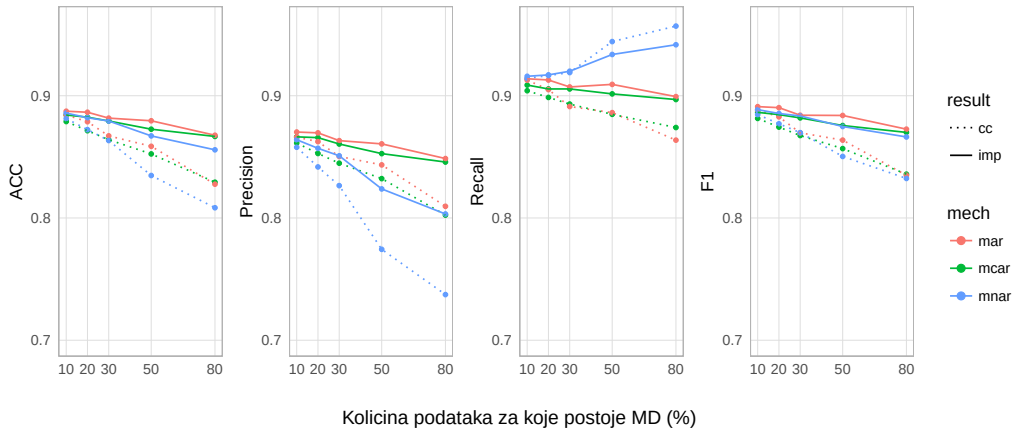
- $mech$: MAR,
- $misPattern$: $mp[p[HDL] = 0, p[TG] = 0, p[GLY] = 0]$,
- $misWeights$: $mw[w(AGE, BMI) = (1, 0.8), w(AGE, BMI) = (1, 0.8),$ $w(AGE, BMI) = (1, 0.8)]$,
- $misOdds$: $mo[o(0, 1, 2, 3), o(3, 2, 1, 0), o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(1, 2), o(2, 1), o(2, 1)]$ za količine od 50% i 80% ,
- $misFreq$: $mf(0.33, 0.33, 0.33)$.

Postavke za scenarije gde MD nastaju po MNAR mehanizmu su:

- $mech$: $MNAR$,
- $misPattern$: $mp[p[HDL] = 0, p[TG] = 0, p[GLY] = 0]$,
- $misWeights$: $mw[w[HDL] = 1, w[TG] = 1, w[GLY] = 1]$,
- $misOdds$: $mo[o(0, 1, 2, 3), o(3, 2, 1, 0), o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(1, 2), o(2, 1), o(2, 1)]$ za količine od 50% i 80% ,

– $misFreq$: $mf(0.33, 0.33, 0.33)$.

Rezultati su navedeni u tabeli A.3 u dodatku, a prikazani na slici 6.7.



Slika 6.7: Scenariji: **MVmd HDL,TG, GLY - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

Opet se može konstatovati da imputacija poboljšava finalni model s tim da je ukupno poboljšanje manje nego kad je u pitanju nedostajanje kod jedne promenljive. Ovo je očekivano s obzirom da je razlika u odnosu na scenarije kad nedostaju vrednosti kod jedne promenljive ta što se za određenu količinu MD sada nedostajanje deli na 3 promenljive pa za svaku pojedinačno imamo i 3 puta manje redova u kojima vrednosti nedostaju. Ovo znači da količina MD za jednu promenljivu ne prelazi 30% zašta smo već u prethodnim slučajevima mogli videti da je okvirna granica gde možemo tolerisati nedostajanje. Ovde bi se potencijalno mogli izvršiti eksperimenti od čak 100% nedostajanja, ali tako da se modeliraju svi paterni koji predstavljaju sve kombinacije ove tri promenljive i dalje analizom diskutovati navedene pretpostavke.

Dalje, ono što je specifično za ovaj slučaj je da sa MNAR mehanizmom, kad nedostaju više laboratorijske vrednosti za promenljivu koja ima MD, dolazi do porasta odziva, a smanjivanja preciznosti. Međutim, ova razlika ne ide u krajnji slučaj gde je za sve pacijente postavljena pozitivna dijagnoza već preciznost ipak zadržava visoke vrednosti. Zato ukupna mera (ACC, F1) zadržava očekivane vrednosti za finalni model. S obzirom na ovakve rezultate, za ovaj konkretan problem, možemo zaključiti da možemo pokušati razvoj finalnog modela nad kompletnim podacima za čak i veće količine MD laboratorijskih vrednosti pod uslovom da se one odnose na vrednosti koje odgovaraju zdravijim osobama.

U svim prethodnim slučajevima i odgovarajućim scenarijima su posmatrane situacije kada vrednosti nedostaju za laboratorijske vrednosti koje predstavljaju spoljašnje pormenljive za izabrani model, odnosno one koje, osim za izračunavanje vrednosti izlazne promenljive, dalje ne učestvuju u formiranju finalnog modela. Opšti zaključak nakon izvršene analize je da imputacija svakako poboljšava finalni prediktivni model u odnosu na korišćenje skupa podataka koji koristi samo kompletne podatke i ova poboljšanja ne zavise značajno od mehanizma nedostojanja, što je u skladu sa ciljevima istraživanja. Kod MNAR mehanizma postoji najviše osetljivosti na promene, ali te promene, u vidu razmene između odziva i preciznosti odgovaraju promenama i kod kompletnog skupa pa se svakako opšta performansa modela poboljšava. Osim ovoga, imamo i zaključak da su dve vrednosti dovoljno informativne i da nema potrebe ukljanjati podatke iz skupa ukoliko jedna od laboratorijskih vrednosti fali.

U nastavku je dat prikaz slučajeva kad nedostaju vrednosti kod ulaznih promenljivih, odnosno kod prediktora.

Scenariji: UVmd - WHtR

Prvo posmatramo slučaj kad se MD javlja kod WHtR vrednosti. S obzirom da se radi o jednostavnom paternu on ima oblik sličan onom koji je prikazan za HDL na slici 6.2.

Postavke za scenarije gde MD nastaju po MCAR mehanizmu su:

- $mech : MCAR$,
- $misPattern : mp[p(WHtR) = 0]$,
- $misWeights : weights.boot$,
- $misOdds : null$,
- $misFreq : mf(1)$.

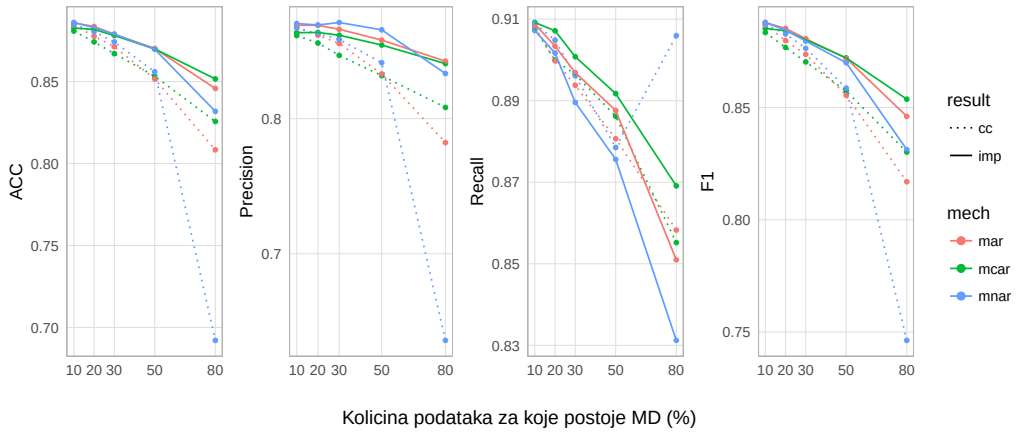
Postavke za scenarije gde MD nastaju po MAR mehanizmu su:

- $mech : MAR$,
- $misPattern : mp[p(WHtR) = 0]$,
- $misWeights : mw[w(AGE, BMI) = (1, 0.8)]$,
- $misOdds : mo[o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1)]$ za količine od 50% i 80%,
- $misFreq : mf(1)$.

Postavke za scenarije gde MD nastaju po MNAR mehanizmu su:

- $mech$: $MNAR$,
- $misPattern$: $mp[p(WHtR) = 0]$,
- $misWeights$: $mw[(w(WHtR) = 1)]$,
- $misOdds$: $mo[o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1)]$ za količine od 50% i 80%,
- $misFreq$: $mf(1)$.

Rezultati su navedeni u tabeli A.4 u dodatku, a prikazani na slici 6.8.



Slika 6.8: Scenariji: **UVmd WHtR - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

Opet možemo primetiti poboljšanje za skup sa imputacijom osim za slučaj kod MNAR mehanizama za veće količine MD. Ako izuzmemo slučaj gde se MD pojavljuje u 80% podataka može se primetiti da skup sa imputacijom zadržava iste trendove kod odziva i preciznosti kao i kompletan skup za sve mehanizme. Iako preciznost zadržava slične vrednosti i pri promeni kolčine, pad u odzivu utiče na opštu tačnost. Primećuje se i da za MNAR mehanizam, kad su veće količine MD većih WHtR vrednosti, dolazi do značajne promene odnosa odziva i preciznosti što ukazuje na povećan broj zdravih pacijenata dijagnostifikovanih kao onih sa sindromom. Međutim, s obzirom na sličnost performansi za sve mehanizme možemo i ovde zaključiti da je imputacija u dobroj meri neosetljiva na mehanizam nedostajanja i generalno omogućava poboljšanje finalnog modela.

Svakako, ukoliko novi skup ima MD kod ove promenljive predlaže se imputacija pre formiranja modela iako je razlika u poboljšanju manja nego kod nedostajanja laboratorijskih vrednosti. Takođe, samo ukoliko je početni skup bio

takav da je uklonjena veća količina većih WHtR vrednosti možemo zaključiti da je to uklanjanje potencijalno predstavljalo pogrešnu odluku.

Scenariji: UVmd - BMI

Naredni scenariji se odnose na vrednosti parametra BMI. Postavke su slične kao u prethodno posmatranim primerima s tim da se za realizaciju MAR mehnizma koristio uzrast pacijenata i njegove WHtR vrednosti.

Postavke za scenarije gde MD nastaju po MCAR mehanizmu su:

- $mech : MCAR$,
- $misPattern : mp[p(BMI) = 0]$,
- $misWeights : weights.boot$,
- $misOdds : null$,
- $misFreq : mf(1)$.

Postavke za scenarije gde MD nastaju po MAR mehanizmu su:

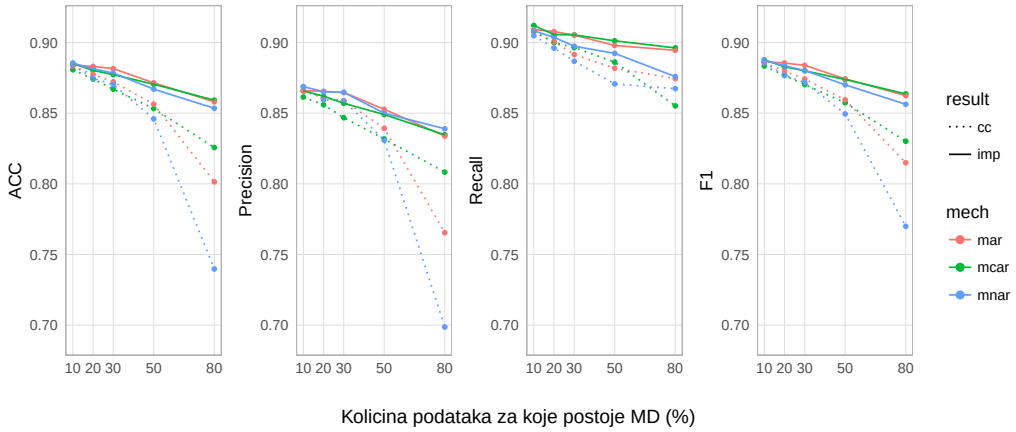
- $mech : MAR$,
- $misPattern : mp[p(BMI) = 0]$,
- $misWeights : mw[w(AGE, WHtR)] = (1, 0.8)]$,
- $misOdds : mo[o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1)]$ za količine od 50% i 80%,
- $misFreq : mf(1)$.

Postavke za scenarije gde MD nastaju po MNAR mehanizmu su:

- $mech : MNAR$,
- $misPattern : mp[p(BMI) = 0]$,
- $misWeights : mw[(w(BMI) = 1]$,
- $misOdds : mo[o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1)]$ za količine od 50% i 80%,
- $misFreq : mf(1)$.

Rezultati su navedeni u tabeli A.5 u dodatku, a prikazani na slici 6.9.

Kao i do sada, utvrđujemo da se imputacijom poboljšavaju performanse modela s tim da se kod BMI vidi znatno poboljšanje preciznosti za sve mehanizme posebno za veće količine MD. Stoga, i u slučaju nedostajanja BMI vrednosti, može se preporučiti imputacija pre formiranja finalnog prediktivnog modela. Takođe,



Slika 6.9: Scenariji: **UVmd BMI - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

kao i u većini prethodnih slučajeva, zaključujemo da eventualno prethodno uklanjanje kompletnih redova je, ukoliko su nedostajale BMI vrednosti, bilo nepotrebno.

Scenariji: MVmd - BMI + WhtR

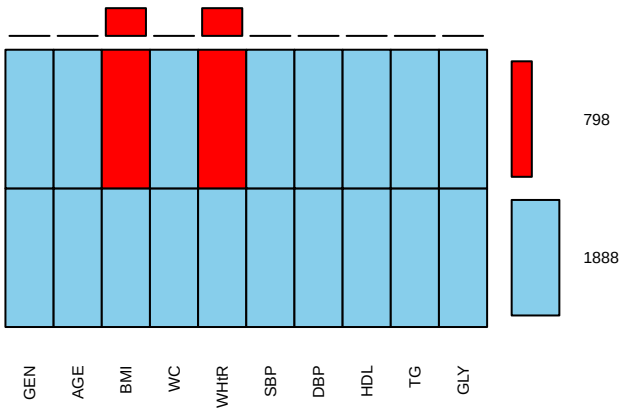
Nakon zasebne analize, kroz naredne scenarije su posmatrana nedostajanja vrednosti za BMI i WHtR s obzirom da su ove dve vrednosti značajne za finalnu dijagnozu, a i očekivano je da u realnim situacijama mogu neodostajati u isto vreme. Za realizaciju MAR mehanizma, ovaj put je uzet u obzir samo uzrast dok je za MNAR mehanizam obema promenljivama data podjednaka važnost. Primer jedne instance paterna za MAR mehanizam i 30% nedostajanja je prikazan na slici 6.10.

Postavke za scenarije gde MD nastaju po MCAR mehanizmu su:

- $mech : MCAR$,
- $misPattern : mp[p(BMI, WHtR) = 0]$,
- $misWeights : weights.boot$,
- $misOdds : null$,
- $misFreq : mf(1)$.

Postavke za scenarije gde MD nastaju po MAR mehanizmu su:

- $mech : MAR$,
- $misPattern : mp[p(BMI, WHtR) = (0, 0)]$,
- $misWeights : mw[w(AGE) = (1)]$,



Slika 6.10: Scenariji: **BMI+WHtR MAR 0.3 Pattern**

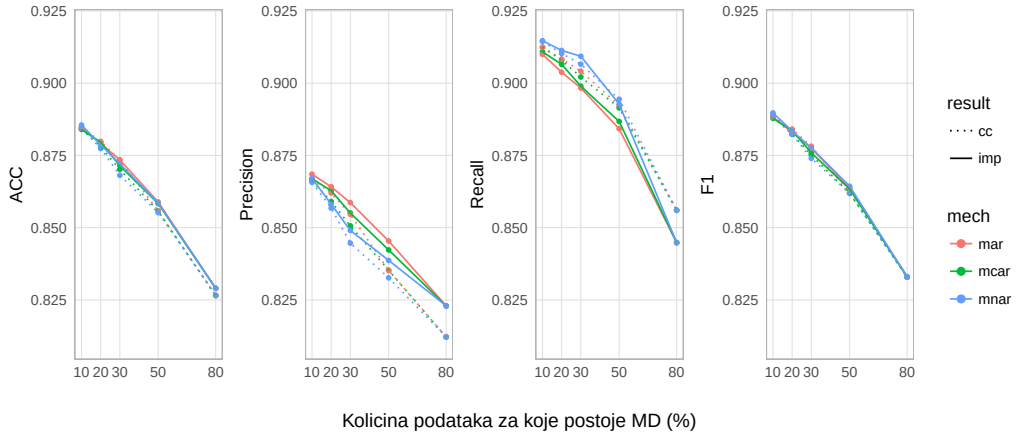
- $misOdds$: $mo[o(3,2,1,0)]$ za količine od 10%, 20% i 30% i $mo[o(2,1)]$ za količine od 50% i 80%,
- $misFreq$: $mf(1)$.

Postavke za scenarije gde MD nastaju po MNAR mehanizmu su:

- $mech$: $MNAR$,
- $misPattern$: $mp[p(BMI, WHtR) = 0]$,
- $misWeights$: $mw[w(BMI, WHtR) = 1]$,
- $misOdds$: $mo[o(3,2,1,0)]$ za količine od 10%, 20% i 30% i $mo[o(2,1)]$ za količine od 50% i 80%,
- $misFreq$: $mf(1)$.

Rezultati su navedeni u tabeli A.6 u dodatku, a prikazani na slici 6.11.

Ovo je primer slučaja gde imputacija ne obezbeđuje bolji finalni model, bar za izabrani i testirani metod. Rezultati i trendovi promena performansi nad imputiranim skupom u potpunosti prate one nad kompletnim skupom podataka. Međutim, ono što je značajno je da iako nema poboljšanja ne postoje ni razlike između samih mehanizama nedostajanja. Ovo jeste u skladu sa pretpostavkom da su ove dve vrednosti jedne od najznačajnijih za predikciju sindroma i ukoliko nedostaju u isto vreme njihove vrednosti se ne mogu rekonstruisati tako da se dobije bolji finalni prediktivni model. To upućuje na zaključak da ukoliko ove vrednosti kod nekih posmatranja fale u isto vreme, iako imputacija neće škoditi, ta posmatranja se mogu isključiti iz skupa za dalje obučavanje finalnog modela. Takođe, ukoliko



Slika 6.11: Scenariji: **MVmd BMI+WHtR - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

su takva posmatranja bila inicijalno isključena iz polaznog skupa onda ja takva odluka o isključivanju i bila neškodljiva. Može se pretpostaviti da bi se i drugi algoritmi imputacije, realizovani metodama mašisnkog učenja, slično ponašali, ali uvek postoji mogućnost zasebne analize i primene drugih metoda posebno za ovaj slučaj.

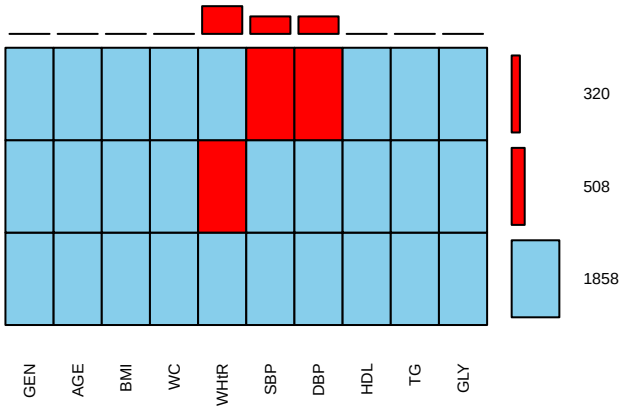
Scenariji: MVmd - WHtR, SBP+DBP

U nastavku su razmatrani složeniji paterni nedostajanja gde se nisu posebno razlikovali scenariji za MCAR, MAR i MNAR mehanizme već su se formirale njihove kombinacije u okviru jednog scenarija.

Prvi slučaj podrazumeva nedostajanje WHtR vrednosti i vrednosti vezanih za krvni pritisak, SBP i DBP. U prvom scenariju je podrazumevano da obe vrednosti za pritisak nedostaju u isto vreme, ali da je to nedostajanje međusobno isključivo sa nedostajanjem WHtR vrednosti s tim da se nedostajanje WHtR nešto češće pojavljuje nego patern u kom nedostaju informacije o pritisku. Sve vrednosti su uklonjene po MAR mehanizmu, u zavisnosti od uzrasta i indeksa telesne mase. Primer realizacije uklanjanja po ovim paternima je prikazan na slici 6.12.

Postavke za ove scenarije su sledeće:

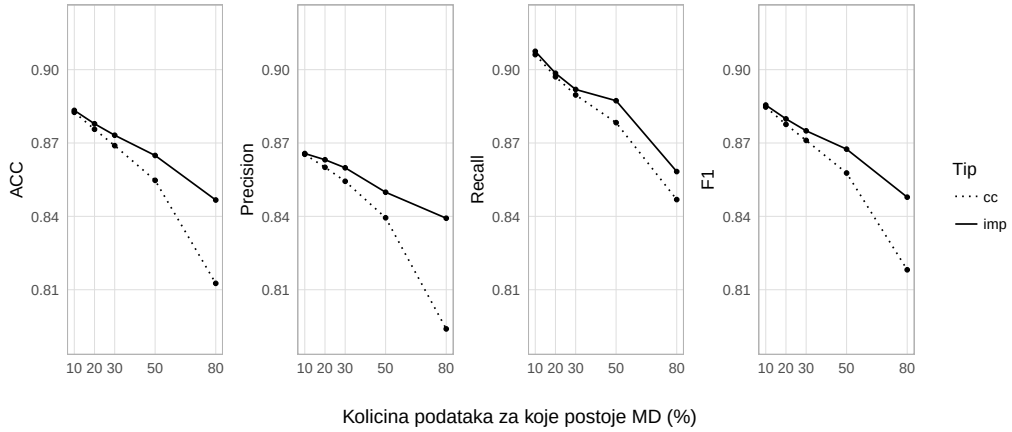
- $mech : MAR$,
- $misPattern : mp[p(WHtR) = 0, p(SBP, DBP) = 0]$,
- $misWeights : mw[w(AGE, BMI) = (1, 0.8), w(AGE, BMI) = (1, 0.8)]$,



Slika 6.12: Scenariji: **WHtR, SBP+DBP MAR 0.3 Pattern**

- $misOdds$: $mo[o(3,2,1,0), o(3,2,1,0)]$ za količine od 10%, 20% i 30% i $mo[o(2,1), o(2,1)]$ za količine od 50% i 80% ,
- $misFreq$: $mf(0.6, 0.4)$.

Rezultati su navedeni u tabeli A.7 u dodatku, a prikazani na slici 6.13.



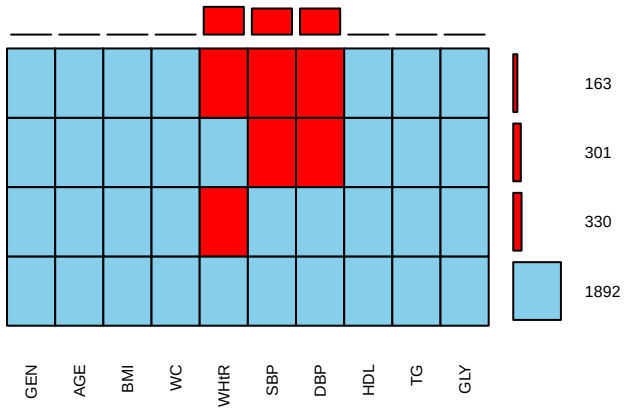
Slika 6.13: Scenariji: **MVmd WHtR,SBP+DBP - MAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

Za navedeni slučaj predložena imputacija poboljšava performanse finalnog prediktivnog modela. Ukolikose rezultati uporede sa prethodnim rezultatima analize pojedinačnih nedostajanja može se utvrditi da oblik promena prati promene uočene i kod slučaja kad vrednosti nedostaju samo kod WHtR (održavanje odziva, poboljšavanje preciznosti) što opet ide u prilog pretpostavci o značaju WHtR pro-

menljive i tome da ne treba vršiti uklanjanje posmatranja gde ove vrednosti nedostaju.

Scenariji: MVmd - WHtR, SBP+DBP, WHtR+SBP+DBP

U prethodnom slučaju su paterni međusobno isključivi odnosno posmatranja imaju MD ili kod WHtR ili kod vrednosti za krvni pritisak. Zato je u sledećem slučaju definisan dodatni patern koji podrazumeva nedostajanje sve tri vrednosti s tim da ovaj patern ima najmanju učestalost pojavljivanja. Primer realizacije ovih paterna je prikazan na slici 6.14.



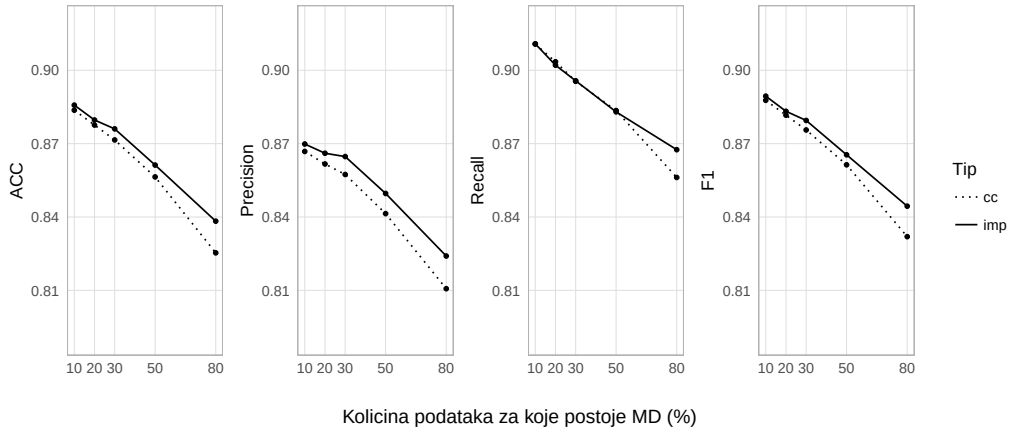
Slika 6.14: Scenariji: **WHtR, SBP+DBP, WHtR+SBP+DBP MAR 0.3 Pattern**

Takođe, za razliku od prethodnog slučaja, uveden je i MNAR mehanizam za WHtR tako da je uslovljeno postojanje veće količine nedostajanja kod manjih vrednosti ove promenljive, ali da je taj uslov manje značajan nego vrednosti za AGE i BMI u poslednjem paternu.

Postavke za ove scenarije su sledeće:

- $mech$: MAR,
- $misPattern$: $mp[p(WHtR) = 0, p(SBP, DBP) = 0, p(WHtR, SBP, DBP) = 0]$,
- $misWeights$: $mw[w(WHtR) = 1), w(AGE, BMI) = (1, 0.8), w(WHtR, AGE, BMI) = (0.8, 1, 1)]$,
- $misOdds$: $mo[o(3, 2, 1, 0), o(3, 2, 1, 0), o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1), o(2, 1), o(2, 1)]$ za količine od 50% i 80%,
- $misFreq$: $mf(0.4, 0.4, 0.2)$.

Rezultati su navedeni u tabeli A.8 u dodatku, a prikazani na slici 6.15

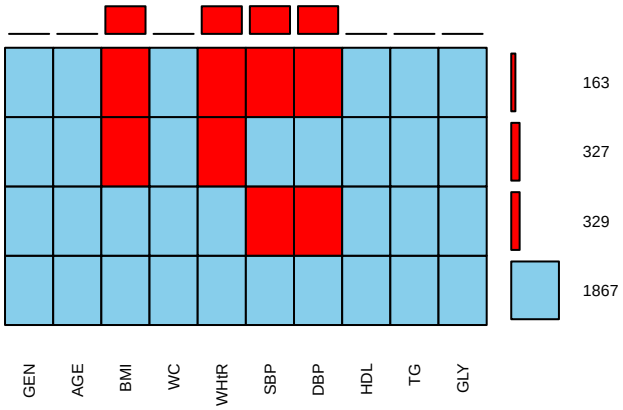


Slika 6.15: Scenariji: **MVmd WHtR,SBP+DBP,WHtR+SBP+DBP**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

Iz rezultata vidimo da imputacija poboljšava performanse finalnog prediktivnog modela, a poboljšanje koje se najviše ogledalo u preciznosti nije toliko izraženo. Na ovom slučaju se može videti kako složenost mehanizama nedostajanja utiče na performanse međutim, ono što je za značajno za ovu analizu je da i pored toga metod imputacije ostaje neosetljiv na mehanizme nedostajanja i uspeva ipak da postigne bolje rezultate.

Scenariji: MVmd - WHtR+BMI, SBP+DBP, WHtR+BMI+SBP+DBP

Na kraju je prikazan najsloženiji slučaj kada je uz promenljive iz prethodnog slučaja dodato i nedostajanje BMI vrednosti gde opet paterni nastaju MAR i MNAR mehanizmima nedostajanja. Primer realizacije ovih paterna je prikazani na slici 6.16.



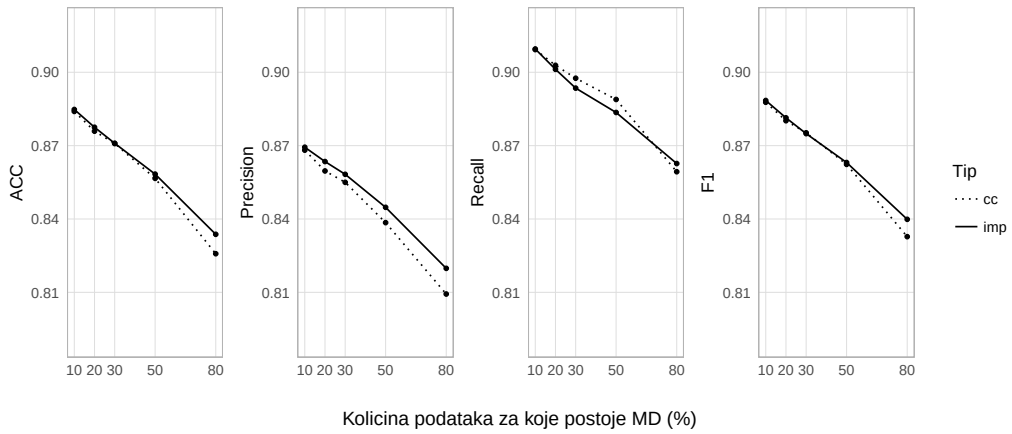
Slika 6.16: Scenariji: **WHtR+BMI, SBP+DBP, WHtR+BMI+SBP+DBP MAR 0.3 Pattern**

Postavke za ove scenarije su sledeće:

- $mech$: MAR,
- $misPattern$: $mp[p(WHtR, BMI) = 0, p(SBP, DBP) = 0, p(WHtR, BMI, SBP, DBP) = 0)$,
- $misWeights$: $mw[w(WHtR, BMI) = 1, w(AGE, BMI) = (1, 1), w(WHtR, BMI) = (1, 1)]$,
- $misOdds$: $mo[o(3, 2, 1, 0), o(3, 2, 1, 0), o(3, 2, 1, 0)]$ za količine od 10%, 20% i 30% i $mo[o(2, 1), o(2, 1), o(2, 1)]$ za količine od 50% i 80%,
- $misFreq$: $mf(0.4, 0.4, 0.2)$.

Rezultati su navedeni u tabeli A.9 u dodatku, a prikazani na slici 6.17.

Kod najsloženijeg slučaja vidimo da imputacija ne poboljšava performanse modela, ali ih ni ne kvari. Nedostatak poboljšanja ne iznenađuje jer podaci koji nedostaju nedostaju za značajne promenljive i to u puno slučajeve istovremeno. Ovakve postavke su razmatrane kao primer ekstremnog slučaja kako bismo utvrdili kako se tada ponaša imputacija. Iako bi se za navedene scenarije u realnosti verovatno odustalo od skupa podataka, kao skupa nad kojm će se razvijati model, utvrđujemo da predložena imputacija ipak uspeva da postigne moguće rezultate što je značajan rezultat.



Slika 6.17: Scenariji: **MVmd WHtR,SBP+DBP,WHtR+SBP+DBP**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije.

6.4.4 Diskusija

Uz pomoć metodologije M2 definisane u 5.4 je izvršena analiza uticaja imputacija različitih slučajeva MD korišćenjem ANN pri razvoju RF modela za predikciju MetS sindroma. Za polazni, kompletan skup su simulacijama uvođene MD u razlčitim količinama, za različite promenljive i koristeći različite mehanizme nedostajanja.

Potencijalno, svaki skup može imati beskonačno scenarija za MD koji se dobijaju kombinacijom različitih parametara od kojih su neki i kontinualani, kao što su težinski faktori koji definišu MAR i MNAR mehanizam. Ovde je analizirana jedna grupa slučajeva nedostajanja koji su izabrani kao predstavnici na osnovu čije analize se može dobiti opštu sliku o ponašanjima finalnih prediktivnih modela modela. Performanse modela su razmatrane uz pomoć opštih mera (tačnost i F1) i kroz analizu promena i nagodba mera koje se odnose na preciznosti i odziv prilikom predikcije na izdvojenim test skupovima.

Za svaki scenario posebno su dati zaključci koji se odnose na ponašanje skupa podataka dobijenog imputacijom u odnosu na korišćenje skupa sa kompletnim podacima prilikom formiranja prediktivnog modela. Na osnovu svega navedenog i opšte slike koja se dobija nakon analize mogu se formirati sledeće zaključci.

Utvrđeno je da bi isključivanje podataka koji sadrže MD kod bilo koje od spoljašnjih, laboratorijskih vrednosti bilo pogrešno jer bi značajno umanjilo performanse finalnog modela za sva tri mehanizma. Što se prediktora tiče, isto važi za BMI I WHtR posebno. Međutim, ako obe vrednosti nedostaju u isto vreme predlo-

žena imputacija ne uspeva da nadoknadi informaciju koja nedostaje pa je eventualno isključivanje ovakvih podataka bilo neškodljivo za finalni model.

Ono što je dalje značajno za istraživanje je da u svim slučajevima predložena imputacija sa ANN uvek omogućava bolje performanse finalnog modela. Koliko će biti poboljšanje i kakav kvalitet će imati zavisi od sledećih faktora: količine MD, značaja neke promenljive za dijagnostiku sindroma, broja pormenljivih koje učestvuju u paternima nedostajanja i same strukture paterna nedostajanja. Pored opšteg poboljšanja performansi, ono što se posebno ističe je da se poboljšanja dešavaju podjedanko za sve mehanizme nedostajanja.

Utvrđen je i značaj analize slučajeva kad se MD javlja kod jedne promenljive čak i kada se ne očekuje pojavljivanje takvog paterna u realnoj situaciji. Takva analiza, pre svega, omogućava precizniju analizu složenijih scenarija jer daje uvide o eventualnim razlozima pojave određenih promena. S druge strane, ona na neki način odgovara i samom procesu izbora značajnih promenljivih koje učestvuju u razvoju finalnog modela. Zato bi bi se već u tom koraku pretporcesiranja, kad se određuju promenljive od značaja, dalo pretpostaviti da li bi eventualna uspešna imputacija imala koristi.

Ovde treba napomenuti da je prilikom realizovane simulacije korišćen veći broj opštih ili generičkih parametara. Za ANN koja se koristila kao metod imputacije, su utvrđeni određeni hiperparametri koji se mogu dodatno optimizovati, a to su: broj skrivenih neurona, slučajna raspodela početnih težina, *limit* i *factor* parametri za *rprop* algoritam. Kao što je navedeno u definiciji metodologije, upotrebljena podešavanja jesu dovoljna za potrebe ovakve vrste simulacije i analize i njihovih ciljeva, ali može se pretpostaviti da, dodatnom optimizacijom svakog od navedenih parametara kao i uvođenjem preciznijeg kriterijuma za zaustavljanje obučavanja ANN, imputacija može ostvariti još bolje rezultate.

7. Zaključak

U ovoj disertaciji je analiziran problem nedostajućih podataka u domenu primene mašinskog učenja pri razvoju prediktivnih modela nad struktuiranim podacima.

Prvo su prikazani okviri, zadaci i potrebe za izvedenim istraživanjem. Opisano je: šta se uopšteno podrazumeva pod problemom nedostajućih podataka, terminologija vezanu za ovaj domen, kako se razlikuju nedostajući podaci po svojoj strukturi i nastajanju i koje metode i procesi se koriste prilikom rada sa ovakvim podacima.

Približen je pojam imputacije podataka i predstavljena studija u kojoj je istraženo kako se veštačke neuralne mreže ponašaju kao metod imputacije laboratorijskih vrednosti u odnosu na druge jednosturke metode imputacije u okviru predikcije kardio-metaboličkog rizika.

Nakon toga je ukazano na značaj analize tretmana nedostajućih podataka pri samoj izgradnji prediktivnih modela. Objašnjeno je na koje načine nedostajući podaci utiču na razvoj prediktivnih modela mašinskog učenja i koje su situacije kad je neophodno razmotriti kako će se postupati u slučajevima kad postoje nedostajući podaci.

Sumiranjem saznanja iz literature i istraživačkih postupaka u domenu problema nedostajućih podataka je definisan predlog metoda za analizu uticaja imputacija namenjih inženjerima iz domena primene mašinskog učenja Metodologija je predstavljena u obliku algoritama koji, zajedno sa napomenama, precizno definišu korake koji obezbeđuju sveobuhvatnu analizu uticaja imputacija pri razvoju prediktivnih modela u odnosu na razvoj modela nad kompletnim podskupovima podataka. U okviru definisanja ovih metodologija skrenuta je pažnja na: moguće probleme koji se mogu javiti, odluke koje treba doneti, modifikacije i optimizacije

koje se odnose na implementaciju predloženih algoritama. Cilj razvoja ove metodologije je bio da se omogući radni okvir i precizan dizajn eksperimenata neophodnih za analizu osetljivosti u odnosu na bilo koju vrstu nedostajućih podataka koristeći procese koji su poznati u domenu mašinskog učenja.

Korišćenjem definisane metodologije je, kroz studiju slučaja, analiziran potencijalni uticaj nedostajućih podataka na konkretnom prediktivnom modelu koji je zasnovan na slučajnim šumama. Tom prilikom su testirani različiti scenariji nedostajanja podataka kroz imputaciju upotrebom veštačkih neuralnih mreža. Eksperimentalno, kroz niz simulacija, je utvrđeno da skupovi nad kojima jeste izvršena imputacija generalno proizvode tačnije finalne modele za sve tipove mehanizma nedostajanja (MCAR, MAR, MNAR) u odnosu na ulazne skupove iz kojih bi nekompletni podaci bili isključeni.

Postignuti su ciljevi istraživanja i svaka od hipoteza, navedenih u uvodu disertacije, je potvrđena. Kroz rezultate ostvarene u okviru istraživanja i prihvaćene za publikaciju u [1] je utvrđeno da se veštačke neuralne mreže mogu iskoristi za jednostruku imputaciju laboratorijskih vrednosti pri razvoju modela predikcije kardio-metaboličkog rizika (X_1). Na osnovu izučavanje dizajna eksperimenata u literaturi i vršenjem empirijskih istraživanja u okviru izrade disertacije je formirana i predložena nova metodologija koja omogućava istraživačima analizu i validaciju izabranih metoda imputacije pri razvoju prediktivnih modela mašinskog učenja (X_2). Kroz studiju slučaja i primenjivanjem ove metodologije je utvrđeno da se veštačke neuralne mreže mogu iskoristiti kao metod imputacije za različite slučajeve nedostajanja podataka pri formiranju modela predikcije metaboličkog sindroma publikovanog u [4]. Metod imputacije preko veštačkih neuralnih mreža se može iskoristiti za različite paterne i količine i sve mehanizme nastajanja nedostajućih nepodataka. U literaturi i dosadašnjem istraživanju predikcije metaboličkog sindroma nije vršena analiza uticaja neke metode imputacije na razvoj finalnog prediktivnog modela na sveobuhvatan način na koji je to izvršeno u okviru istraživanja ove disertacije (X_3).

Osim navedenih rezultata dobijeni su i dodatni doprinosi. Prvo, iako je studija slučaja bila izvršena na jednom konkretnom skupu iz domena medicinske dijagnostike sam proces je opšteg karaktera i ne koristi domeska saznanja pa se ekvivalentno može primeniti u bilo kom domenu, što potvrđuje namenu razvijene i predstavljene metodologije. Drugo, u okviru analize veštačke neualne mreže kao metode imputacije posebno je izdvojena imputacija spoljašnjih promenljivih koje se koriste za izračunavanje izlaznih vrednosti, ali se same ne koriste u izgradnji

finalnog modela što predstavlja dodatni doprinos ukupnom istraživanju s obzirom da su ovakvi scenariji retko razmatrani i prikazani u literaturi.

Na osnovu svega iznetog može se doneti nekoliko zaključaka. Prvo, i pre svega, podaci koji nedostaju se, prilikom razvoja prediktivnih modela u domenu mašinskog učenja, ne smeju olako zanemariti odlukom da se oni isključe iz daljeg istraživanja Drugo, nije nužno i obavezno poznavati detaljno oblast analize nedostajućih podataka u smislu fundamentalne teorije koja se odnosi na njih i koja je definisana u oblasti statistike. Međutim, potrebno je izvršiti određenu analizu osetljivosti uticaja ovakvih podataka za dalji razvoj modela, a veoma je korisno, uz osnovne rezultate o razvoju ili primeni nekog modela mašinskog učenja, prikazati i informacije koje se odnose na nedostajuće podatke prilikom publikovanja prediktivnog modela. Ukoliko je nedostajućih podataka bilo u polaznom skupu treba opisati njihovu strukturu, tretman koji je korišćen i eventualne uticaje na buduću primenu objavljenog modela. Ako u polaznom skupu nije bilo podataka koji nedostaju, a ima osnove da oni mogu postojati u realnoj situaciji, treba navesti ta zapažnja, a idealno i izvršiti analizu uticaja na formiranje prediktivnog modela.

Na kraju, ovo istraživanje ostavlja prostora za sumnju da se i nekim drugim procesima u različitim fazama razvoja prediktivnih modela, u oblasti primene mašinskog učenja, pristupa sa manje pažnje, kao što je to često slučaj sa analizom i tretmanom nedostajućih podataka u fazi pretprocesiranja. Osim preglednih istraživanja koja se tiču pristupa određenim fazama razvoja i primene prediktivnih modela mašinskog učenja, za nastavk rada u ovoj oblasti se otvara i niz drugih istraživačkih pitanja:

- Kako se dalje mogu optimizovati procesi predložene metodologije u odnosu na složenost izvršavanja?
- Da li se mogu implementirati ili modifikovati pojedinačni algoritmi mašinskog učenja koji u sam svoj proces obučavanja ugrađuju tretman nedostajućih podataka na smislen način koristeći, recimo, ideju višestruke imputacije, nenadgledanog učenja ili primenjenje matematike?
- Da li se može kreirati vizualizacija opisanih procesa i rezultata s obzirom da su dostupne vizualizacije u oblasti analize nedostajućih podataka opet usmerene na istraživanje statističkih osobina i procena?
- Kako se dosadašnja saznanja mogu povezati i iskoristiti za detekciju i tretman pogrešnih podataka koji su u svoj suštini nedostajući ali, potencijalno, nose više informacija?

I na kraju, ako se osvrnemo opet na početak i Flekov pojam "misli kolektiva", možemo se zapitati da li je sve dosadašnje istraživanje u domenu problema nedostajućih podataka uslovljeno tom čvrsto formiranom misli i ideji o nedostajućim podacima koja je nastala u statistici, znatno ranije pre široke primene modernih saznanja i tehnologija. Smelo se može pretpostaviti da ako bismo se svesno udaljili od takve, duboko usvojene, terminologije koja podrazumeva MCAR, MAR i MNAR mehanizme, možda bismo mogli doći do potpuno novih naučnih teorija koje se tiču problema nedostajućih podataka, a čije su osnovne ideje zasnovane, ovaj put, na primeni naučnih saznanja iz domena mašinskog učenja ili moderne primenjene matematike.

Literatura

[1] Dunja Vrbaški, Aleksandar Kupusinac, Rade Doroslovački, Edita Stokić, and Dragan Ivetić. Missing data imputation in cardiometabolic risk assessment: A solution based on artificial neural networks. *Computer Science and Information Systems*, 2020. forthcoming.

[2] Roderick J. Little, Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, Joseph W. Hogan, Geert Molenberghs, Susan A. Murphy, James D. Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung J. Shih, Jay P. Siegel, and Hal Stern. The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[3] Cátia M. Salgado, Carlos Azevedo, Hugo Proença, and Susana M. Vieira. *Missing Data*, pages 143–162. Springer International Publishing, Cham, 2016.

[4] Dunja Vrbaški, Milan Vrbaški, Aleksandar Kupusinac, Darko Ivanović, Edita Stokić, Dragan Ivetić, and Ksenija Doroslovački. Methods for algorithmic diagnosis of metabolic syndrome. *Artificial Intelligence in Medicine*, 101:101708, 2019.

[5] Thomas Samuel Kuhn and Staniša Novaković. *Struktura naučnih revolucija*. Nolit, 1974.

[6] Ludwik Fleck. *Genesis and Development of a Scientific Fact, trans.* F. Bradley & Trenn, TJ University of Chicago Press, Chicago, 1979.

[7] Wojciech Sady. Ludwik fleck. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition, 2019. URL https://plato.stanford.edu/archives/win2019/entries/fleck/.

[8] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[9] Akbar K Waljee, Peter DR Higgins, and Amit G Singal. A primer on predictive models. *Clinical and translational gastroenterology*, 5(1):e44, 2014.

[10] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[11] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

[12] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley, 2002.

[13] Grigorios Papageorgiou, Stuart W Grant, Johanna J M Takkenberg, and Mostafa M Mokhles. Statistical primer: how to deal with missing data in scientific research? *Interactive CardioVascular and Thoracic Surgery*, 27(2): 153–158, 05 2018. ISSN 1569-9285.

[14] Rolf H.H. Groenwold, Ian R. White, A. Rogier T. Donders, James R. Carpenter, Douglas G. Altman, and Karel G.M. Moons. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ*, 184(11):1265–1269, 2012.

[15] Douglas G Altman and J Martin Bland. Missing data. *Bmj*, 334(7590):424–424, 2007.

[16] James H. Ware, David Harrington, David J. Hunter, and Ralph B. D'Agostino. Missing data. *New England Journal of Medicine*, 367(14):1353–1354, 2012.

[17] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[18] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147–77, 2002.

[19] John W. Graham, Patricio E. Cumsille, and Elvira Elek-Fisk. Methods for Handling Missing Data. In *Handbook of Psychology*, pages 87–114. John Wiley & Sons, Inc., 2003.

[20] Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83 (404):1198–1202, 1988.

[21] Rianne Margaretha Schouten and Gerko Vink. The dance of the mechanisms: How observed information influences the validity of missingness assumptions. *Sociological Methods & Research*, 2018.

[22] Billingsley Kaambwa, Stirling Bryan, and Lucinda Billingham. Do the methods used to analyse missing data really matter? an examination of data from an observational study of intermediate care patients. *BMC research notes*, 5 (1):330, 2012.

[23] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[24] Andrea Marshall, Douglas G Altman, Patrick Royston, and Roger L Holder. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology*, 10(1):7, 2010.

[25] AMG Ali, SJ Dawson, FM Blows, E Provenzano, IO Ellis, Laura Baglietto, D Huntsman, C Caldas, and PD Pharoah. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. *British journal of cancer*, 104(4):693–699, 2011.

[26] Marianne Riksheim Stavseth, Thomas Clausen, and Jo Røislien. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE open medicine*, 7, 2019.

[27] Katya L Masconi, Tandi E Matsha, Rajiv T Erasmus, and Andre P Kengne. Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of south africa. *PloS one*, 10(9), 2015.

[28] Jungyeon Choi, Olaf M Dekkers, and Saskia le Cessie. A comparison of different methods to handle missing data in the context of propensity score analysis. *European journal of epidemiology*, 34(1):23–36, 2019.

[29] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

[30] Rachael A Hughes, Jon Heron, Jonathan AC Sterne, and Kate Tilling. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International journal of epidemiology*, 1:11, 2019.

[31] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[33] Erkki Pesonen, Matti Eskelinen, and Martti Juhola. Treatment of missing data values in a neural network based decision support system for acute abdominal pain. *Artificial Intelligence in Medicine*, 13(3):139 – 146, 1998.

[34] Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, Manuel López-Coello, and María-Dolores Cubiles de-la Vega. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, 24(1): 121 – 129, 2011.

[35] Brett K Beaulieu-Jones, Jason H Moore, and Consortium. Missing data imputation in the electronic health record using deeply learned autocencoders. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22: 207–218, 2016.

[36] Collins Leke, Tshilidzi Marwala, and Satyakama Paul. Proposition of a Theoretical Model for Missing Data Imputation using Deep Learning and Evolutionary Algorithms. *arXiv*, 2015.

[37] Collins Leke and Tshilidzi Marwala. Missing Data Estimation in High-Dimensional Datasets: A Swarm Intelligence-Deep Neural Network Approach. pages 259–270. Springer, Cham, 2016.

[38] Jaisheel Mistry, Fulufhelo V Nelwamondo, and Tshilidzi Marwala. Missing data estimation using principle component analysis and autoassociative neural networks. *Journal of Systemics, Cybernatics and Informatics*, 7(3):72–79, 2009.

[39] Mussa Abdella and Tshilidzi Marwala. The use of genetic algorithms and neural networks to approximate missing data in database. In *IEEE 3rd International Conference on Computational Cybernetics, 2005. ICCC 2005.*, pages 207–212. IEEE, 2005.

[40] Vukosi N Marivate, Fulufhelo V Nelwamondo, and Tshilidzi Marwala. Investigation into the use of autoencoder neural networks, principal component analysis and support vector regression in estimating missing hiv data. *IFAC Proceedings Volumes*, 41(2):682–689, 2008.

[41] Yanjie Duan, Yisheng Lv, Yu-Liang Liu, and Fei-Yue Wang. An efficient realization of deep learning for traffic data imputation. *Transportation Research Part C: Emerging Technologies*, 72:168 – 181, 2016.

[42] Jung-Woo Kim and Yakov A. Pachepsky. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for swat streamflow simulation. *Journal of Hydrology*, 394(3):305 – 314, 2010.

[43] Nicholas J Tierney, Fiona A Harden, Maurice J Harden, and Kerrie L Mengersen. Using decision trees to understand structure in missing data. *BMJ open*, 5(6):e007450, 2015.

[44] Talayeh Razzaghi, Oleg Roderick, Ilya Safro, and Nicholas Marko. Multilevel weighted support vector machine for classification on healthcare data with missing values. *PloS one*, 11(5), 2016.

[45] Marek Smieja, Łukasz Struski, Jacek Tabor, Bartosz Zieliński, and Przemysław Spurek. Processing of missing data by neural networks, 2018.

[46] Aleksandar Kupusinac, Rade Doroslovački, Dušan Malbaški, Biljana Srdić, and Edita Stokić. A primary estimation of the cardiometabolic risk by using artificial neural networks. *Computers in Biology and Medicine*, 43(6):751–757, 2013.

[47] Christian Igel, Marc Toussaint, Wan Weishui, József Szabados, and Marcel G. de Bruin. Rprop using the natural gradient. In *Trends and Applications in Constructive Approximation*, pages 259–272. Birkhäuser Basel, 2005.

[48] Martin Riedmiller. Advanced supervised learning in multi-layer perceptrons — from backpropagation to adaptive learning algorithms. *Computer Standards and Interfaces*, 16(3):265 – 278, 1994. ISSN 0920-5489.

[49] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2018. URL https://CRAN.R-project.org/package=caret. R package version 6.0-80.

[50] Pedro M Domingos. A few useful things to know about machine learning. *Commun. acm*, 55(10):78–87, 2012.

[51] Phil Gibbs and Sugihara Hiroshi. What is occam's razor?, 1997. URL http://math.ucr.edu/home/baez/physics/General/occam.html. accessed: 2019-11.

[52] Stef van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2012.

[53] Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

[54] Spyros Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, 1993.

[55] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[56] Bradley Efron and Robert J Tibshirani. *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*. Division of Biostatistics, Stanford University, 1995.

[57] Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382): 316–331, 1983.

[58] Rahul C. Deo. Machine Learning in Medicine. *Circulation*, 132:1920–1930, 2015.

[59] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *CoRR*, 2015.

[60] Riccardo Miotto, Li Li, Brian A. Kidd, Joel T. Dudley, and P. Agarwal. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6, 2016.

[61] K. R. Foster, R. Koprowski, and J. D. Skufca. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering OnLine*, 13(1):94, Jul 2014.

[62] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17, 2009.

[63] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[64] Scott M. Grundy, James I. Cleeman, Stephen R. Daniels, Karen A. Donato, Robert H. Eckel, Barry A. Franklin, David J. Gordon, Ronald M. Krauss, Peter J. Savage, Sidney C. Smith, John A. Spertus, and Fernando Costa. Diagnosis and Management of the Metabolic Syndrome. *Circulation*, 112:2735–2752, 2005.

[65] Hana Rosolova and Barbora Nussbaumerova. Cardio-metabolic risk prediction should be superior to cardiovascular risk assessment in primary prevention of cardiovascular diseases. *The EPMA journal*, 2:15–26, 2011.

[66] Karel GM Moons, Patrick Royston, Yvonne Vergouwe, Diederick E Grobbee, and Douglas G Altman. Prognosis and prognostic research: what, why, and how? *Bmj*, 338:b375, 2009.

[67] Katya L Masconi, Tandi E Matsha, Justin B Echouffo-Tcheugui, Rajiv T Erasmus, and Andre P Kengne. Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. *The EPMA Journal*, 2015.

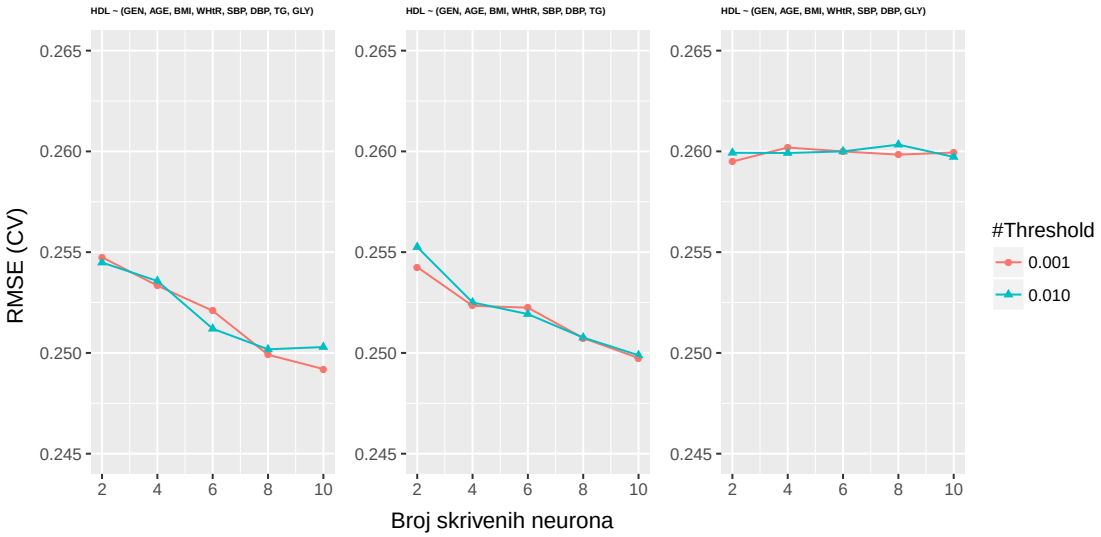
[68] Stefan Fritsch and Frauke Guenther. *neuralnet: Training of Neural Networks*, 2016. URL https://CRAN.R-project.org/package=neuralnet. R package version 1.33.

[69] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

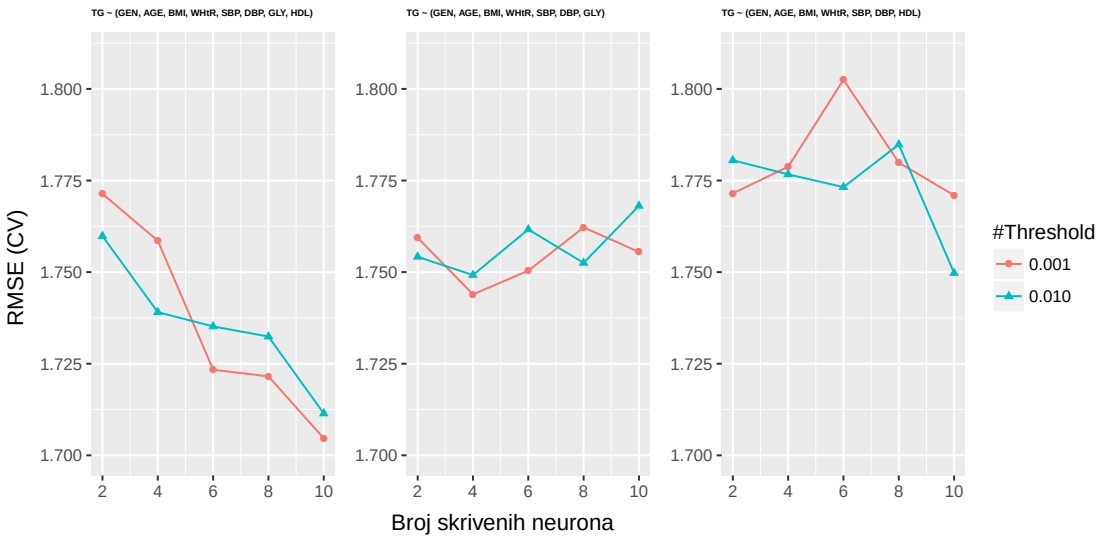
[70] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

A. Dodatak

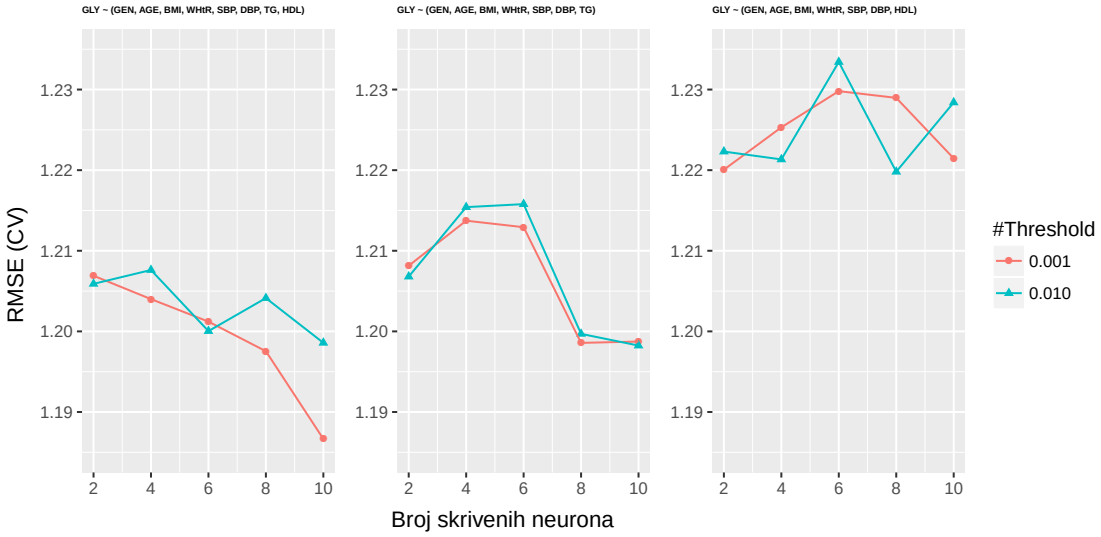
A.1 Studija slučaja: Eksperimentalni rezultati pripreme ANN modela za imputaciju



Slika A.1: Određivanje broja neurona i threshold vrednosti za neuralne mreže neophodne za imputaciju HDL za različite strukture mreže



Slika A.2: Određivanje broja neurona i threshold vrednosti za neuralne mreže neophodne za imputaciju TG za različite strukture mreže



Slika A.3: Određivanje broja neurona i threshold vrednosti za neuralne mreže neophodne za imputaciju GLY za različite strukture mreže

Promenljiva	Ulaz za ANN	HN	Threshold	RMSE
HDL	p + GLY, TG	2	0.001	0.255
HDL	p + GLY, TG	2	0.010	0.254
HDL	p + GLY, TG	4	0.001	0.253
HDL	p + GLY, TG	4	0.010	0.254
HDL	p + GLY, TG	6	0.001	0.252
HDL	p + GLY, TG	6	0.010	0.251
HDL	p + GLY, TG	8	0.001	0.250
HDL	p + GLY, TG	8	0.010	0.250
HDL	p + GLY, TG	10	0.001	0.249
HDL	p + GLY, TG	10	0.010	0.250
HDL	p + GLY	2	0.001	0.259
HDL	p + GLY	2	0.010	0.260
HDL	p + GLY	4	0.001	0.260
HDL	p + GLY	4	0.010	0.260
HDL	p + GLY	6	0.001	0.260
HDL	p + GLY	6	0.010	0.260
HDL	p + GLY	8	0.001	0.260
HDL	p + GLY	8	0.010	0.260
HDL	p + GLY	10	0.001	0.260
HDL	p + GLY	10	0.010	0.260
HDL	p + TG	2	0.001	0.254
HDL	p + TG	2	0.010	0.255
HDL	p + TG	4	0.001	0.252
HDL	p + TG	4	0.010	0.253
HDL	p + TG	6	0.001	0.252
HDL	p + TG	6	0.010	0.252
HDL	p + TG	8	0.001	0.251
HDL	p + TG	8	0.010	0.251
HDL	p + TG	10	0.001	0.250
HDL	p + TG	10	0.010	0.250
TG	p + GLY, HDL	2	0.001	1.771
TG	p + GLY, HDL	2	0.010	1.760
TG	p + GLY, HDL	4	0.001	1.759
TG	p + GLY, HDL	4	0.010	1.739
TG	p + GLY, HDL	6	0.001	1.723
TG	p + GLY, HDL	6	0.010	1.735
TG	p + GLY, HDL	8	0.001	1.722

TG	p + GLY, HDL	8	0.010	1.732
TG	p + GLY, HDL	10	0.001	1.705
TG	p + GLY, HDL	10	0.010	1.711
TG	p + GLY	2	0.001	1.759
TG	p + GLY	2	0.010	1.754
TG	p + GLY	4	0.001	1.744
TG	p + GLY	4	0.010	1.749
TG	p + GLY	6	0.001	1.750
TG	p + GLY	6	0.010	1.762
TG	p + GLY	8	0.001	1.762
TG	p + GLY	8	0.010	1.753
TG	p + GLY	10	0.001	1.756
TG	p + GLY	10	0.010	1.768
TG	p + HDL	2	0.001	1.772
TG	p + HDL	2	0.010	1.781
TG	p + HDL	4	0.001	1.779
TG	p + HDL	4	0.010	1.777
TG	p + HDL	6	0.001	1.803
TG	p + HDL	6	0.010	1.773
TG	p + HDL	8	0.001	1.780
TG	p + HDL	8	0.010	1.785
TG	p + HDL	10	0.001	1.771
TG	p + HDL	10	0.010	1.750
GLY	p + TG, HDL	2	0.001	1.207
GLY	p + TG, HDL	2	0.010	1.206
GLY	p + TG, HDL	4	0.001	1.204
GLY	p + TG, HDL	4	0.010	1.208
GLY	p + TG, HDL	6	0.001	1.201
GLY	p + TG, HDL	6	0.010	1.200
GLY	p + TG, HDL	8	0.001	1.198
GLY	p + TG, HDL	8	0.010	1.204
GLY	p + TG, HDL	10	0.001	1.187
GLY	p + TG, HDL	10	0.010	1.199
GLY	p + TG	2	0.001	1.208
GLY	p + TG	2	0.010	1.207
GLY	p + TG	4	0.001	1.214
GLY	p + TG	4	0.010	1.215
GLY	p + TG	6	0.001	1.213

GLY	p + TG	6	0.010	1.216
GLY	p + TG	8	0.001	1.199
GLY	p + TG	8	0.010	1.200
GLY	p + TG	10	0.001	1.199
GLY	p + TG	10	0.010	1.198
GLY	p + HDL	2	0.001	1.220
GLY	p + HDL	2	0.010	1.222
GLY	p + HDL	4	0.001	1.225
GLY	p + HDL	4	0.010	1.221
GLY	p + HDL	6	0.001	1.230
GLY	p + HDL	6	0.010	1.233
GLY	p + HDL	8	0.001	1.229
GLY	p + HDL	8	0.010	1.220
GLY	p + HDL	10	0.001	1.221
GLY	p + HDL	10	0.010	1.228

Tabela A.1: Određivanje broja neurona i threshold vrednosti za neuralne mreže neophodne za imputaciju. Sa p je označen skup prediktora za MetS model. Za svaku promenljivu su prikazane prosečne vrednosti RMSE za svaku kombinaciju parametara i strukture mreže

Listing A.1: Izračunavanje skora za procenu postojanja metaboličkog sindroma

```
calculateMets <- function(person)
{
   mets = 0

   if ((person["BMI"] > 30) ||
       (person["GEN"] == 2 && person["WC"] >= 80) ||
       (person["GEN"] == 1 && person["WC"] >= 94))
   {
      p = 0
      if(person["SBP"] >= 130 || person["DBP"] >= 85) { p = p + 1 }
      if ((person["GEN"] == 2 && person["HDL"] < 1.29) ||
         (person["GEN"] == 1 && person["HDL"] < 1.03)) { p = p + 1 }
      if (person["TG"] >= 1.7) { p = p + 1 }
      if (person["GLY"] >= 5.6) { p = p + 1 }
      if (p >= 2) { mets = 1; }
   }
   return (mets)
}
```

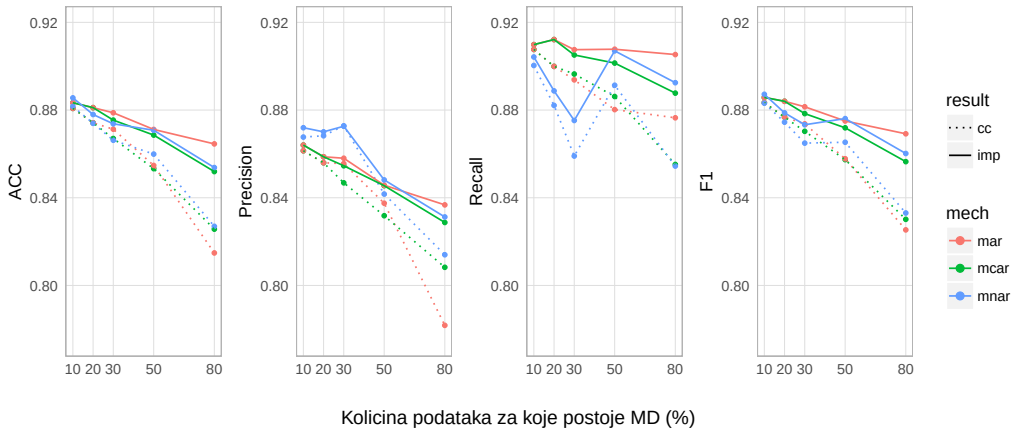
A.2 Studija slučaja: Eksperimentalni rezultati imputacija

VAR	MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
HDL	mcar	10	cc	0.881	0.761	0.861	0.908	0.856	0.883
HDL	mcar	10	imp	0.884	0.768	0.866	0.910	0.860	0.887
HDL	mcar	20	cc	0.874	0.748	0.856	0.900	0.850	0.877
HDL	mcar	20	imp	0.883	0.765	0.865	0.908	0.859	0.885
HDL	mcar	30	cc	0.867	0.733	0.847	0.896	0.839	0.870
HDL	mcar	30	imp	0.878	0.756	0.859	0.905	0.854	0.881
HDL	mcar	50	cc	0.853	0.706	0.832	0.886	0.822	0.857
HDL	mcar	50	imp	0.874	0.746	0.855	0.900	0.849	0.876
HDL	mcar	80	cc	0.826	0.651	0.808	0.855	0.798	0.830
HDL	mcar	80	imp	0.861	0.721	0.845	0.884	0.840	0.864
HDL	mar	10	cc	0.884	0.768	0.867	0.908	0.862	0.887
HDL	mar	10	imp	0.888	0.775	0.869	0.914	0.864	0.890
HDL	mar	20	cc	0.878	0.755	0.862	0.900	0.857	0.880
HDL	mar	20	imp	0.885	0.770	0.867	0.911	0.861	0.888
HDL	mar	30	cc	0.871	0.742	0.856	0.894	0.851	0.874
HDL	mar	30	imp	0.882	0.764	0.864	0.909	0.858	0.885
HDL	mar	50	cc	0.855	0.708	0.838	0.880	0.830	0.858
HDL	mar	50	imp	0.876	0.751	0.858	0.902	0.852	0.878
HDL	mar	80	cc	0.815	0.628	0.782	0.876	0.754	0.825
HDL	mar	80	imp	0.874	0.746	0.856	0.898	0.850	0.876
HDL	mnar	10	cc	0.883	0.765	0.871	0.900	0.868	0.884
HDL	mnar	10	imp	0.887	0.773	0.873	0.906	0.869	0.888
HDL	mnar	20	cc	0.878	0.756	0.875	0.884	0.875	0.878
HDL	mnar	20	imp	0.884	0.768	0.873	0.900	0.870	0.886
HDL	mnar	30	cc	0.872	0.743	0.875	0.869	0.877	0.871
HDL	mnar	30	imp	0.879	0.757	0.874	0.886	0.874	0.879
HDL	mnar	50	cc	0.857	0.713	0.838	0.889	0.824	0.862
HDL	mnar	50	imp	0.874	0.748	0.858	0.902	0.848	0.879
HDL	mnar	80	cc	0.826	0.651	0.811	0.856	0.796	0.833
HDL	mnar	80	imp	0.861	0.721	0.847	0.887	0.836	0.866
TG	mcar	10	cc	0.881	0.761	0.861	0.908	0.856	0.883
TG	mcar	10	imp	0.883	0.766	0.864	0.910	0.859	0.886
TG	mcar	20	cc	0.874	0.748	0.856	0.900	0.850	0.877

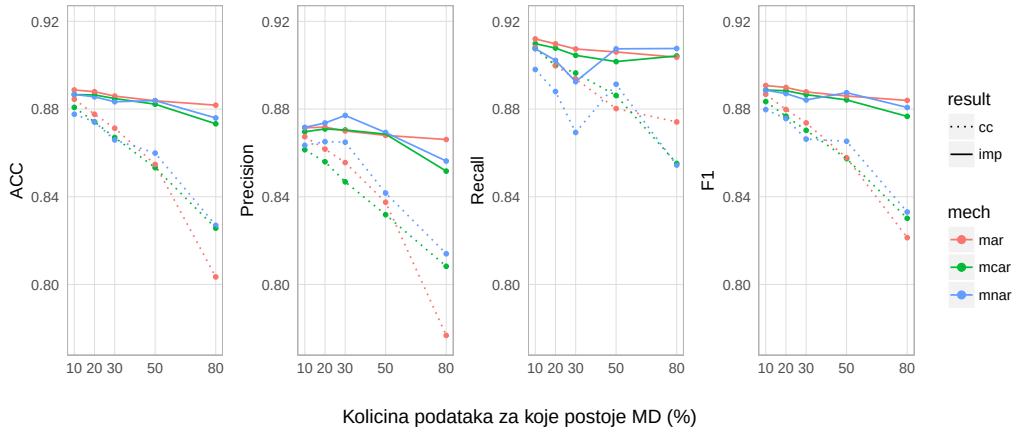
TG	mcar	20	imp	0.881	0.762	0.859	0.912	0.852	0.884
TG	mcar	30	cc	0.867	0.733	0.847	0.896	0.839	0.870
TG	mcar	30	imp	0.875	0.750	0.855	0.905	0.848	0.878
TG	mcar	50	cc	0.853	0.706	0.832	0.886	0.822	0.857
TG	mcar	50	imp	0.869	0.736	0.846	0.901	0.837	0.872
TG	mcar	80	cc	0.826	0.651	0.808	0.855	0.798	0.830
TG	mcar	80	imp	0.852	0.703	0.829	0.888	0.818	0.856
TG	mar	10	cc	0.881	0.761	0.861	0.908	0.856	0.883
TG	mar	10	imp	0.883	0.766	0.864	0.910	0.859	0.886
TG	mar	20	cc	0.874	0.748	0.856	0.900	0.850	0.877
TG	mar	20	imp	0.881	0.762	0.859	0.912	0.852	0.884
TG	mar	30	cc	0.871	0.742	0.856	0.894	0.851	0.874
TG	mar	30	imp	0.879	0.757	0.858	0.908	0.852	0.882
TG	mar	50	cc	0.855	0.708	0.838	0.880	0.830	0.858
TG	mar	50	imp	0.871	0.742	0.846	0.908	0.836	0.875
TG	mar	80	cc	0.815	0.628	0.782	0.876	0.754	0.825
TG	mar	80	imp	0.865	0.728	0.837	0.905	0.825	0.869
TG	mnar	10	cc	0.882	0.763	0.868	0.900	0.865	0.883
TG	mnar	10	imp	0.886	0.770	0.872	0.904	0.869	0.887
TG	mnar	20	cc	0.874	0.747	0.868	0.882	0.868	0.874
TG	mnar	20	imp	0.878	0.755	0.870	0.889	0.869	0.879
TG	mnar	30	cc	0.866	0.732	0.873	0.859	0.876	0.865
TG	mnar	30	imp	0.874	0.747	0.873	0.875	0.874	0.873
TG	mnar	50	cc	0.860	0.719	0.842	0.891	0.828	0.865
TG	mnar	50	imp	0.871	0.740	0.848	0.907	0.834	0.876
TG	mnar	80	cc	0.827	0.653	0.814	0.854	0.800	0.833
TG	mnar	80	imp	0.854	0.706	0.831	0.892	0.814	0.860
GLY	mcar	10	cc	0.881	0.761	0.861	0.908	0.856	0.883
GLY	mcar	10	imp	0.887	0.772	0.870	0.910	0.865	0.889
GLY	mcar	20	cc	0.874	0.748	0.856	0.900	0.850	0.877
GLY	mcar	20	imp	0.886	0.772	0.871	0.908	0.867	0.888
GLY	mcar	30	cc	0.867	0.733	0.847	0.896	0.839	0.870
GLY	mcar	30	imp	0.885	0.769	0.870	0.905	0.867	0.887
GLY	mcar	50	cc	0.853	0.706	0.832	0.886	0.822	0.857
GLY	mcar	50	imp	0.882	0.764	0.869	0.902	0.865	0.884
GLY	mcar	80	cc	0.826	0.651	0.808	0.855	0.798	0.830
GLY	mcar	80	imp	0.873	0.746	0.852	0.904	0.844	0.877
GLY	mar	10	cc	0.884	0.768	0.867	0.908	0.862	0.887

GLY	mar	10	imp	0.889	0.777	0.871	0.912	0.867	0.891
GLY	mar	20	cc	0.878	0.755	0.862	0.900	0.857	0.880
GLY	mar	20	imp	0.888	0.775	0.872	0.910	0.868	0.890
GLY	mar	30	cc	0.871	0.742	0.856	0.894	0.851	0.874
GLY	mar	30	imp	0.886	0.771	0.870	0.907	0.866	0.888
GLY	mar	50	cc	0.855	0.708	0.838	0.880	0.830	0.858
GLY	mar	50	imp	0.884	0.767	0.868	0.906	0.864	0.886
GLY	mar	80	cc	0.803	0.604	0.777	0.874	0.729	0.821
GLY	mar	80	imp	0.882	0.763	0.866	0.904	0.861	0.884
GLY	mnar	10	cc	0.878	0.755	0.863	0.898	0.859	0.880
GLY	mnar	10	imp	0.887	0.773	0.872	0.908	0.868	0.889
GLY	mnar	20	cc	0.874	0.748	0.865	0.888	0.863	0.876
GLY	mnar	20	imp	0.886	0.770	0.874	0.902	0.871	0.887
GLY	mnar	30	cc	0.866	0.731	0.865	0.869	0.865	0.866
GLY	mnar	30	imp	0.883	0.766	0.877	0.893	0.876	0.884
GLY	mnar	50	cc	0.860	0.719	0.842	0.891	0.828	0.865
GLY	mnar	50	imp	0.884	0.767	0.869	0.907	0.860	0.887
GLY	mnar	80	cc	0.827	0.653	0.814	0.854	0.800	0.833
GLY	mnar	80	imp	0.876	0.751	0.856	0.908	0.844	0.881

Tabela A.2: Scenariji: **UVmd HDL/TG/GLY + MCAR/MAR/MNAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije



Slika A.4: Scenariji: **UVmd TG - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije



Slika A.5: Scenariji: **UVmd GLY - MCAR/MAR/MNAR**. Prikaz preformansi finalnog prediktivnog modela za sve scenarije

MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
mcar	10	cc	0.879	0.757	0.861	0.904	0.856	0.881
mcar	10	imp	0.884	0.768	0.866	0.909	0.862	0.886
mcar	20	cc	0.871	0.742	0.853	0.899	0.846	0.874
mcar	20	imp	0.882	0.764	0.866	0.906	0.861	0.885
mcar	30	cc	0.864	0.727	0.845	0.893	0.837	0.867
mcar	30	imp	0.879	0.758	0.861	0.906	0.855	0.882
mcar	50	cc	0.852	0.704	0.832	0.885	0.822	0.857
mcar	50	imp	0.873	0.745	0.853	0.901	0.846	0.876
mcar	80	cc	0.829	0.658	0.802	0.874	0.786	0.836
mcar	80	imp	0.867	0.733	0.846	0.897	0.838	0.870
mar	10	cc	0.884	0.768	0.866	0.913	0.856	0.888
mar	10	imp	0.887	0.774	0.870	0.914	0.861	0.891
mar	20	cc	0.879	0.756	0.862	0.905	0.853	0.883
mar	20	imp	0.886	0.772	0.870	0.913	0.860	0.890
mar	30	cc	0.867	0.734	0.850	0.891	0.845	0.870
mar	30	imp	0.882	0.763	0.863	0.907	0.858	0.884
mar	50	cc	0.859	0.716	0.843	0.886	0.832	0.864
mar	50	imp	0.880	0.758	0.861	0.909	0.849	0.884
mar	80	cc	0.828	0.654	0.810	0.864	0.792	0.835
mar	80	imp	0.868	0.734	0.849	0.899	0.836	0.873
mnar	10	cc	0.881	0.762	0.858	0.914	0.850	0.885
mnar	10	imp	0.886	0.771	0.864	0.916	0.858	0.889
mnar	20	cc	0.872	0.744	0.842	0.917	0.829	0.877
mnar	20	imp	0.882	0.763	0.857	0.917	0.849	0.885
mnar	30	cc	0.863	0.726	0.826	0.919	0.808	0.870
mnar	30	imp	0.879	0.758	0.851	0.920	0.840	0.883
mnar	50	cc	0.835	0.669	0.774	0.944	0.726	0.850
mnar	50	imp	0.867	0.733	0.824	0.934	0.801	0.875
mnar	80	cc	0.808	0.616	0.737	0.957	0.659	0.832
mnar	80	imp	0.856	0.711	0.803	0.942	0.772	0.866

Tabela A.3: Scenariji: **MVmd HDL, TG, GY - MCAR/MAR/MNAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

MECH	MP	TYPE	ACC	Kappa	Precison	Recall	Spec	F1
mcar	10	cc	0.881	0.761	0.861	0.908	0.856	0.883
mcar	10	imp	0.883	0.765	0.864	0.909	0.858	0.885
mcar	20	cc	0.874	0.748	0.856	0.900	0.850	0.877
mcar	20	imp	0.882	0.763	0.864	0.907	0.859	0.884
mcar	30	cc	0.867	0.733	0.847	0.896	0.839	0.870
mcar	30	imp	0.878	0.756	0.862	0.901	0.858	0.880
mcar	50	cc	0.853	0.706	0.832	0.886	0.822	0.857
mcar	50	imp	0.870	0.739	0.854	0.892	0.850	0.872
mcar	80	cc	0.826	0.651	0.808	0.855	0.798	0.830
mcar	80	imp	0.852	0.703	0.841	0.869	0.837	0.854
mar	10	cc	0.884	0.768	0.867	0.908	0.862	0.887
mar	10	imp	0.886	0.771	0.869	0.909	0.865	0.888
mar	20	cc	0.878	0.755	0.862	0.900	0.857	0.880
mar	20	imp	0.884	0.767	0.869	0.903	0.866	0.885
mar	30	cc	0.871	0.742	0.856	0.894	0.851	0.874
mar	30	imp	0.879	0.757	0.866	0.897	0.863	0.881
mar	50	cc	0.852	0.702	0.833	0.881	0.825	0.855
mar	50	imp	0.870	0.740	0.858	0.888	0.855	0.872
mar	80	cc	0.808	0.616	0.782	0.858	0.760	0.817
mar	80	imp	0.846	0.691	0.842	0.851	0.843	0.846
mnar	10	cc	0.885	0.769	0.867	0.909	0.863	0.887
mnar	10	imp	0.886	0.771	0.870	0.907	0.867	0.888
mnar	20	cc	0.880	0.760	0.863	0.905	0.858	0.883
mnar	20	imp	0.883	0.765	0.869	0.902	0.867	0.884
mnar	30	cc	0.874	0.748	0.859	0.896	0.854	0.876
mnar	30	imp	0.879	0.757	0.871	0.890	0.870	0.880
mnar	50	cc	0.856	0.711	0.841	0.879	0.836	0.859
mnar	50	imp	0.870	0.739	0.866	0.876	0.866	0.870
mnar	80	cc	0.692	0.381	0.636	0.906	0.476	0.746
mnar	80	imp	0.832	0.663	0.833	0.831	0.834	0.831

Tabela A.4: Scenariji: **UVmd WHtR - MCAR/MAR/MNAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
mcar	10	cc	0.881	0.761	0.861	0.908	0.856	0.883
mcar	10	imp	0.885	0.770	0.866	0.912	0.860	0.888
mcar	20	cc	0.874	0.748	0.856	0.900	0.850	0.877
mcar	20	imp	0.880	0.760	0.862	0.906	0.856	0.883
mcar	30	cc	0.867	0.733	0.847	0.896	0.839	0.870
mcar	30	imp	0.877	0.754	0.857	0.905	0.851	0.880
mcar	50	cc	0.853	0.706	0.832	0.886	0.822	0.857
mcar	50	imp	0.870	0.740	0.849	0.901	0.841	0.874
mcar	80	cc	0.826	0.651	0.808	0.855	0.798	0.830
mcar	80	imp	0.859	0.718	0.835	0.896	0.824	0.864
mar	10	cc	0.883	0.766	0.866	0.908	0.861	0.886
mar	10	imp	0.884	0.767	0.866	0.909	0.861	0.886
mar	20	cc	0.878	0.754	0.861	0.901	0.856	0.880
mar	20	imp	0.883	0.766	0.865	0.908	0.861	0.886
mar	30	cc	0.872	0.744	0.859	0.891	0.855	0.874
mar	30	imp	0.882	0.762	0.865	0.905	0.860	0.884
mar	50	cc	0.856	0.712	0.839	0.882	0.832	0.859
mar	50	imp	0.872	0.742	0.853	0.898	0.846	0.874
mar	80	cc	0.801	0.602	0.765	0.874	0.730	0.815
mar	80	imp	0.858	0.715	0.834	0.895	0.823	0.862
mnar	10	cc	0.884	0.767	0.869	0.905	0.865	0.886
mnar	10	imp	0.885	0.770	0.869	0.908	0.864	0.888
mnar	20	cc	0.875	0.749	0.859	0.896	0.855	0.877
mnar	20	imp	0.881	0.762	0.865	0.904	0.861	0.884
mnar	30	cc	0.870	0.740	0.859	0.887	0.856	0.872
mnar	30	imp	0.878	0.756	0.865	0.897	0.861	0.880
mnar	50	cc	0.846	0.691	0.831	0.871	0.823	0.850
mnar	50	imp	0.867	0.733	0.850	0.892	0.843	0.870
mnar	80	cc	0.740	0.480	0.699	0.867	0.615	0.770
mnar	80	imp	0.854	0.706	0.839	0.876	0.833	0.856

Tabela A.5: Scenariji: **UVmd BMI - MCAR/MAR/MNAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
mcar	10	cc	0.884	0.767	0.866	0.912	0.856	0.888
mcar	10	imp	0.884	0.767	0.867	0.911	0.858	0.888
mcar	20	cc	0.878	0.755	0.859	0.908	0.848	0.882
mcar	20	imp	0.880	0.758	0.863	0.907	0.853	0.884
mcar	30	cc	0.870	0.739	0.851	0.902	0.838	0.875
mcar	30	imp	0.871	0.742	0.855	0.899	0.844	0.876
mcar	50	cc	0.856	0.710	0.835	0.891	0.820	0.862
mcar	50	imp	0.858	0.716	0.842	0.887	0.830	0.863
mcar	80	cc	0.827	0.652	0.812	0.856	0.797	0.833
mcar	80	imp	0.829	0.657	0.823	0.845	0.814	0.833
mar	10	cc	0.885	0.768	0.867	0.912	0.857	0.888
mar	10	imp	0.885	0.768	0.869	0.910	0.859	0.888
mar	20	cc	0.880	0.759	0.862	0.908	0.852	0.884
mar	20	imp	0.879	0.757	0.864	0.904	0.855	0.883
mar	30	cc	0.873	0.746	0.855	0.904	0.843	0.878
mar	30	imp	0.873	0.746	0.859	0.898	0.849	0.877
mar	50	cc	0.856	0.711	0.835	0.892	0.820	0.862
mar	50	imp	0.859	0.717	0.845	0.884	0.835	0.864
mar	80	cc	0.827	0.652	0.812	0.856	0.797	0.833
mar	80	imp	0.829	0.657	0.823	0.845	0.814	0.833
mnar	10	cc	0.885	0.768	0.866	0.914	0.855	0.889
mnar	10	imp	0.886	0.770	0.867	0.915	0.857	0.890
mnar	20	cc	0.877	0.754	0.857	0.910	0.845	0.882
mnar	20	imp	0.879	0.756	0.858	0.911	0.846	0.883
mnar	30	cc	0.868	0.735	0.845	0.907	0.830	0.874
mnar	30	imp	0.872	0.743	0.849	0.909	0.835	0.878
mnar	50	cc	0.855	0.709	0.833	0.894	0.816	0.862
mnar	50	imp	0.858	0.716	0.839	0.893	0.824	0.864
mnar	80	cc	0.827	0.652	0.812	0.856	0.797	0.833
mnar	80	imp	0.829	0.657	0.823	0.845	0.814	0.833

Tabela A.6: Scenariji: **MVmd BMI+WHtR - MCAR/MAR/MNAR**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
mar	10	cc	0.883	0.764	0.865	0.906	0.861	0.885
mar	10	imp	0.883	0.766	0.866	0.908	0.861	0.886
mar	20	cc	0.876	0.751	0.860	0.897	0.856	0.878
mar	20	imp	0.878	0.755	0.863	0.899	0.859	0.880
mar	30	cc	0.869	0.737	0.854	0.890	0.850	0.871
mar	30	imp	0.873	0.746	0.860	0.892	0.856	0.875
mar	50	cc	0.855	0.709	0.839	0.878	0.833	0.858
mar	50	imp	0.865	0.729	0.850	0.887	0.845	0.867
mar	80	cc	0.813	0.624	0.794	0.847	0.779	0.818
mar	80	imp	0.847	0.693	0.839	0.858	0.837	0.848

Tabela A.7: Scenariji: **MVmd WHtR,SBP + DBP**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

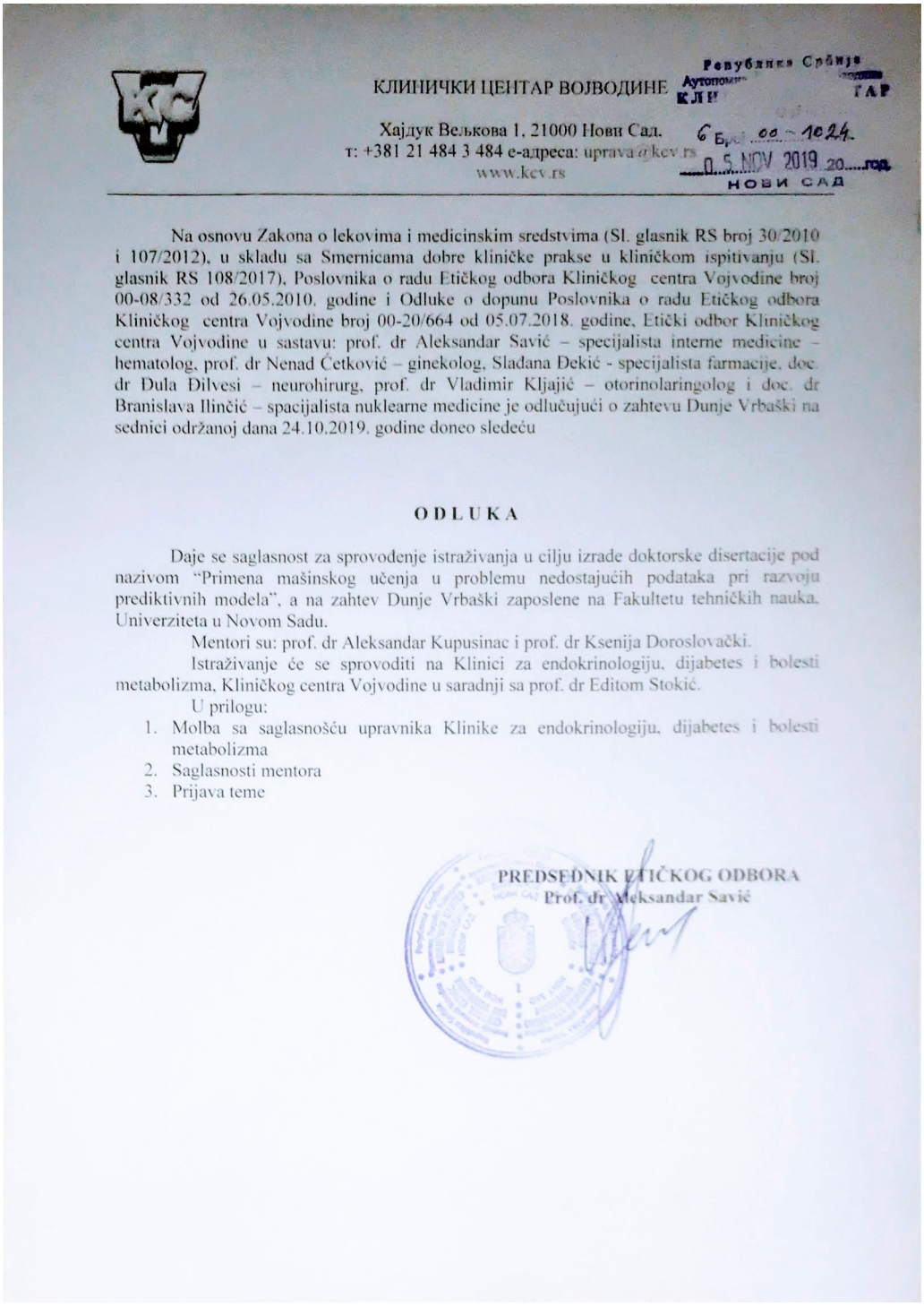
MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
mar/mnar	10	cc	0.884	0.766	0.867	0.911	0.857	0.888
mar/mnar	10	imp	0.886	0.771	0.870	0.911	0.861	0.889
mar/mnar	20	cc	0.878	0.754	0.862	0.903	0.852	0.882
mar/mnar	20	imp	0.880	0.758	0.866	0.902	0.858	0.883
mar/mnar	30	cc	0.872	0.742	0.857	0.896	0.848	0.876
mar/mnar	30	imp	0.876	0.751	0.865	0.896	0.857	0.879
mar/mnar	50	cc	0.856	0.712	0.841	0.884	0.830	0.861
mar/mnar	50	imp	0.861	0.722	0.850	0.883	0.840	0.865
mar/mnar	80	cc	0.825	0.650	0.811	0.856	0.795	0.832
mar/mnar	80	imp	0.838	0.676	0.824	0.868	0.810	0.844

Tabela A.8: Scenariji: **MVmd WHtR,SBP+DBP,WHtR+SBP+DBP**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

MECH	MP	TYPE	ACC	Kappa	Precison	Recall	SPec	F1
mar/mnar	10	cc	0.884	0.767	0.868	0.909	0.859	0.888
mar/mnar	10	imp	0.885	0.769	0.869	0.910	0.861	0.888
mar/mnar	20	cc	0.876	0.751	0.860	0.903	0.849	0.880
mar/mnar	20	imp	0.878	0.754	0.863	0.901	0.854	0.881
mar/mnar	30	cc	0.871	0.741	0.855	0.898	0.845	0.875
mar/mnar	30	imp	0.871	0.741	0.858	0.894	0.849	0.875
mar/mnar	50	cc	0.857	0.712	0.839	0.889	0.825	0.862
mar/mnar	50	imp	0.858	0.716	0.845	0.884	0.834	0.863
mar/mnar	80	cc	0.826	0.650	0.809	0.859	0.793	0.833
mar/mnar	80	imp	0.834	0.667	0.820	0.863	0.806	0.840

Tabela A.9: Scenariji: **MVmd WHtR+BMI,SBP+DBP,WHtR+BMI+SBP+DBP**. Upoređivanje performansi finalnog modela razvijanog samo nad kompletnim podacima (cc) i nad svim podacima uz imputaciju MD (imp). Prikazuje prosečne rezultate ostvarene kroz simulacije.

A.3 Saglasnost etičkog odbora Kliničkog centra Vojvodine za sprovođenje istraživanja



КЛИНИЧКИ ЦЕНТАР ВОЈВОДИНЕ

Хајдук Вељкова 1, 21000 Нови Сад.
т: +381 21 484 3 484 e-адреса: uprava@kcv.rs
www.kcv.rs

Република Србија
Аутономна покрајина
КЛИНИЧКИ ЦЕНТАР
НОВИ САД

Број 00-1024
05 NOV 2019 20... год.

Na osnovu Zakona o lekovima i medicinskim sredstvima (Sl. glasnik RS broj 30/2010 i 107/2012), u skladu sa Smernicama dobre kliničke prakse u kliničkom ispitivanju (Sl. glasnik RS 108/2017), Poslovnika o radu Etičkog odbora Kliničkog centra Vojvodine broj 00-08/332 od 26.05.2010. godine i Odluke o dopunu Poslovnika o radu Etičkog odbora Kliničkog centra Vojvodine broj 00-20/664 od 05.07.2018. godine, Etički odbor Kliničkog centra Vojvodine u sastavu: prof. dr Aleksandar Savić – specijalista interne medicine – hematolog, prof. dr Nenad Četković – ginekolog, Sladana Đekić - specijalista farmacije, doc. dr Đula Đilvesi – neurohirurg, prof. dr Vladimir Kljajić – otorinolaringolog i doc. dr Branislava Ilinčić – spacijalista nuklearne medicine je odlučujući o zahtevu Dunje Vrbaški na sednici održanoj dana 24.10.2019. godine doneo sledeću

ODLUKA

Daje se saglasnost za sprovođenje istraživanja u cilju izrade doktorske disertacije pod nazivom "Primena mašinskog učenja u problemu nedostajućih podataka pri razvoju prediktivnih modela", a na zahtev Dunje Vrbaški zaposlene na Fakultetu tehničkih nauka, Univerziteta u Novom Sadu.

Mentori su: prof. dr Aleksandar Kupusinac i prof. dr Ksenija Doroslovački.

Istraživanje će se sprovoditi na Klinici za endokrinologiju, dijabetes i bolesti metabolizma, Kliničkog centra Vojvodine u saradnji sa prof. dr Editom Stokić.

U prilogu:

1. Molba sa saglasnošću upravnika Klinike za endokrinologiju, dijabetes i bolesti metabolizma
2. Saglasnosti mentora
3. Prijava teme

PREDSEDNIK ETIČKOG ODBORA
Prof. dr Aleksandar Savić

Slika A.6: Odluka Etičkog odbora za davanje saglasnosti za sprovođenje istraživanja na Klinici za endokrinologiju, dijabetes i bolesti metabolizma