Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.8
DOI: https://doi.org/10.24867/13BE10Petrovic

STRATEGIJE ZA RAD SA NEDOSTAJUĆIM VREDNOSTIMA

STRATEGIES FOR DEALING WITH MISSING VALUES

Sreten Petrović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *Ovaj rad se bavi predstavljanjem strategija za rad sa nedostajućim vrijednostima i pokazivanjem njihovih prednosti, mana i efikasnosti u kombinaciji sa algoritmima mašinskog učenja prilikom predviđanja popularnosti mobilnih aplikacija.*

Ključne reči: *strategije, nedostajuće vrijednosti, algoritmi, mašinsko učenje, predikcija, mobilne aplikacije*

Abstract – *This paper deals with the presentation of strategies for dealing with missing values and showing their advantages, disadvantages and efficiency in combination with machine learning algorithms while predicting the popularity of mobile applications.*

Keywords: *strategies, missing values, algorithms, machine learning, prediction, mobile applications*

1. UVOD

Cilj rada jeste definisanje i podjela strategija za rad sa nedostajućim vrijednostima i navođenje primjera korišćenja istih kako bi se riješio konkretan problem koji će se, takođe, opisati u radu. Izvršiće se podjela nedostajućih vrijednosti na određene tipove sa kratkim pregledom osobina i funkcionalnosti. Upotrebom različitih metoda i tehnika za analizu, razumijevanje i sređivanje skupa podataka, formiraće se skup podataka nad kojim će se primjenjivati opisane strategije za rad sa nedostajućim vrijednostima. Kratak pregled algoritama mašinskog učenja, koji će se koristiti u kombinaciji sa strategijama, će biti predstavljen. Pregled implementacije, analiza dobijenih rezultata i diskusija o efikasnosti strategija i algoritama prilikom rješavanja problema biće navedena kako bi se pojasnilo koja kombinacija strategije i algoritma je bila najuspješnija prilikom rješavanja.

2. TEORIJSKI PREGLED PROBLEMA

Problem koji će se analizirati jeste na koji način strategije za rad sa nedostajućim vrijednostima utiču na konkretan projekat, da li je i na koji način primjena strategija dovela do poboljšanja rezultata. Strategije će se primjenjivati za formiranje skupa podataka, a taj skup podataka će služiti kao osnova za predviđanje popularnosti mobilnih aplikacija. Popularnost mobilnih aplikacija će se vršiti na osnovu određenih obilježja aplikacija koja se nalaze u okviru skupa podataka. Prije primjene samih strategija neophodno je prikupiti, analizirati, razumjeti i provjeriti podatke, kako bi se što bolje definisalo koji podaci su odgovarajući, a koji ne i kako bi se započeo proces njihove obrade.

Organizacija podataka u okviru skupa podataka predstavlja važan preduslov za dalje čišćenje, uklanjanje ili imputaciju nedostajućih podataka. Uzroci zbog čega se vrše navedene manipulacije nad podacima su nedostajuće vrijednosti, pogrešno unijeti podaci, kao i domenski pogrešni podaci.

Prema Rubinu [1] postoji nekoliko tipova nedostajućih vrijednosti podataka:

1. Potpuno slučajno - ***Missing completely at random*** (***MCAR***) - predstavlja vjerovatnoću da nedostajući podaci nisu povezani sa određenom vrijednošću koja bi trebala da se dobije, kao ni sa vrijednostima ostalih obilježja
2. Slučajno - ***Missing at random*** (***MAR***) - za razliku od ***MCAR***-a ovaj tip podataka ima vezu sa vrijednostima ostalih obilježja
3. Neslučajno - ***Missing not at random*** (***MNAR***) - ovaj tip podataka je najproblematičniji i jedini način da se dobiju nepristrasne procjene nedostajućih podataka jeste da se napravi odgovarajući model koji bi ih obradio korišćenjem metoda za predviđanje nedostajućih vrijednosti

Rješavanje ovog problema omogućava da se korisniku obezbijedi brz i kvalitetan pristup aplikacijama koje pripadaju kategorijama najpopularnijih aplikacija i koje se poklapaju sa interesnom sferom korisnika.

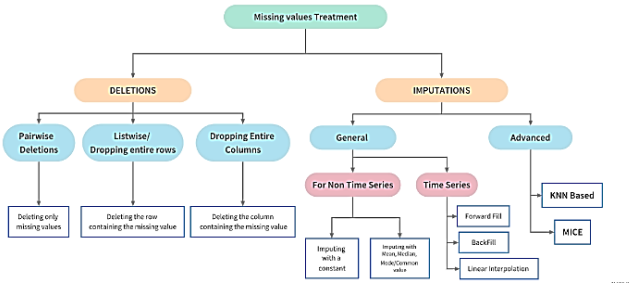
3. METODOLOGIJA

Strategije za rad sa nedostajućim vrijednostima predstavljaju proces pretvaranja i mapiranja neobrađenih podataka u drugu formu sa ciljem da se podaci učine pogodnim za dalju upotrebu i istraživanje. Podjela strategija na određeni broj faza i podfaza, kao i izučavanje i detaljno opisivanje svake pojedinačno, izvršeno je na osnovu literature [3-5].

Strategije za rad sa nedostajućim vrijednostima možemo podijeliti na dvije osnovne grupe, a to su brisanje i imputacija podataka, koje se dalje dijele u odgovarajuće podgrupe.

U nastavku je dat kratak pregled strategija za rad sa nedostajućim vrijednostima, a sumirani pregled prikazan je na slici 1, koja je preuzeta iz literature [2].

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kupusinac, red. prof.



Slika 1. *Dijagram podele strategija za rad sa nedostajućim vrijednostima*

3.1 Brisanje podataka

Predstavlja najjednostavniju strategiju za rad sa nedostajućim vrijednostima. Brisanje podataka se može podijeliti na tri strategije, a to su:

1. Brisanje cijelog opažaja (***Listwise deletion***) - strategija za brisanje cijelog reda skupa podatka ukoliko postoji nedostajuća vrijednost u okviru tog reda
2. Brisanje nedostajuće vrijednosti u okviru opažaja (***Pairwise deletion***) - na osnovu korelacionih matrica može se izmjeriti veza između dva obilježja; za svaki par obilježja za koji su podaci dostupni, koeficijent korelacije će te podatke uzeti u obzir
3. Brisanje cijele kolone (***Dropping entire column***) – vrši brisanje kolone ukoliko postoje nedostajuće vrijednosti podataka u kolonama

3.2 Imputacija

Cilj imputacije jeste upotreba poznatih odnosa u postojećim vrijednostima skupa podataka koji bi se koristili u procjeni nedostajućih vrijednosti. Imputacija sprečava gubitak podataka i na taj način se poboljšava tačnost skupova podataka. Može se podijeliti u dvije grupe, a to su opšte i napredne imputacije. Opšta grupa imputacija se dalje dijeli na ***Non Time Series*** i ***Time Series***. ***Non time series*** se dijeli na ***mean***, ***median*** i ***mode*** imputacije i na imputaciju sa konstantnom vrijednošću. ***Time series*** se dijeli na ***Last Observation Carried Forward*** (***LOCF***), ***Next Observation Carried Backward*** (***NOCB***) i linearnu interpolaciju. Naprednoj grupi imputacija pripadaju ***k*** najbližih komšija (***k nearest neighbors – k-NN***) i višestruka imputacija korišćenjem ulančanih jednačina (***Multiple imputation by chained equations - MICE***).

Mean, ***median*** i ***mode*** imputacije imaju za cilj da nedostajuće vrijednosti u okviru jednog obilježja zamijene vrijednošću dobijene na osnovu poznatih vrijednosti tog obilježja, ne uzimajući u obzir odnos sa drugim vrijednostima iz drugih obilježja. Nedostajuća vrijednost se korišćenjem ***mean*** imputacije dobija kao srednja vrijednost poznatih vrijednosti u okviru obilježja u kojoj se nalazi nedostajuća vrijednost.

Za razliku od ***mean*** imputacije, ***median*** imputacija predstavlja zamjenu nedostajućih podataka sa vrijednošću koja se nalazi u sredini sortiranog skupa. U slučaju da je ukupan broj podataka paran broj, ***median*** se računa kao srednja vrijednost dvije vrijednosti koje se nalaze u sredini sortiranog skupa. ***Mode*** imputacija predstavlja zamjenu nedostajuće vrijednosti podatka sa vrijednošću podatka koji se najčešće pojavljuje u jednom obilježju.

Upotrebom imputacije sa konstantnom vrijednošću se nedostajuća vrijednost zamjenjuje sa nepromjenjivom vrijednošću koja se određuje na osnovu tipa i vrijednosti podataka iz obilježja.

Last Observation Carried Forward **(LOCF) i** ***Next Observation Carried Backward*** **(NOCB)** se koriste ukoliko su u pitanju podaci vremenskih serija. Podaci vremenskih serija predstavljaju niz vrijednosti podataka koji su dobijeni uzastopno, često sa jednakim intervalima između njih. Razlika između ***LOCF*** i ***NOCB*** imputacije jeste ta što se uz pomoć ***LOCF*** imputacije svaka nedostajuća vrijednost mijenja sa posljednjom posmatranom vrijednošću, dok se sa ***NOCB*** svaka nedostajuća vrijednost mijenja sa prvom posmatranom vrijednošću, ali unazad.

Linearna interpolacija predstavlja linearni odnos između podataka i koristi vrijednosti koje ne nedostaju iz susjednih podataka za izračunavanje nedostajuće vrijednosti. Interpolacija predstavlja matematičku metodu koja prilagođava funkciju podacima i koristi je za ekstrapolaciju podataka koji nedostaju.

k-NN imputacija koristi algoritam ***k*** najbližih komšija za ocjenjivanje i zamjenu nedostajućih vrijednosti. Prilikom implementacije ***k-NN*** ne mora se pripremiti prediktivni model za svako obilježje sa nedostajućim vrijednostima. Nedostatak ovog algoritma predstavlja mala brzina izvršavanja. Nije uvijek lako odrediti ***k*** i mjeru sličnosti koja će se koristiti. Osjetljiv je na obilježja koji nisu od značaja.

Višestruka imputacija korišćenjem ulančanih jednačina (***MICE***) može da obrađuje promjenljive različitih tipova. Prvi zadatak prilikom implementacije ove strategije je utvrđivanje koje promjenljive će biti uključene u proces imputacije. U ***MICE*** proceduri pokreće se niz regresionih modela pri čemu se svaka promjenljiva sa podacima koji nedostaju modeluje uslovno prema ostalim promjenljivim u podacima. Generalno se izvodi 10 ciklusa, ali je potrebno istraživanje kako bi se identifikovao optimalan broj ciklusa prilikom unosa podataka pod različitim uslovima. Ideja je da do kraja ciklusa raspodela parametara koji upravljaju imputacijama konvergiraju u smislu da postane stabilna. Kada se završi naznačeni broj ciklusa, čitav postupak imputiranja se ponavlja da bi se generisali višestruki imputirani skupovi podataka. Posmatrani podaci će biti isti u imputiranim skupovima podataka, razlikovaće se samo vrijednosti koje su nedostajale.

Metoda klasifikacije je metoda potpornih vektora (***SVM***) koja predstavlja algoritam koji se može koristiti i za klasifikaciju i za regresiju. Obično se koristi kada su podaci numerički i kada imaju samo dvije klase. Ukoliko je potrebno može se proširiti na više klasa i moguć je rad i sa nenumeričkim podacima. Cilj algoritma je da na osnovu hiper-ravni (***hyperplane***) koja razdvaja podatke kreira model uz pomoć koga će se predvidjeti kojoj klasi pripada nedostajući podatak. Ukoliko su podaci linearno razdvojeni, potrebno je pronaći maksimalnu marginu koja u odnosu na hiper-ravan razdvaja podatke. Zatim se na osnovu hiper-ravni, odnosno, njene jednačine kreira model. Na osnovu modela se računa rastojanje podataka

od hiper-ravni i na osnovu dobijenog rezultata se određuje kojoj klasi pripada. Margina predstavlja rastojanje od hiper-ravni do potpornih vektora, dok potporni vektori predstavljaju podatke koji su najbliže hiper-ravni i na taj način potporni vektori utiču na položaj ravni. Korišćenjem potpornih vektora se pronalazi maksimalna margina hiper-ravne. Potporni vektori pomažu u izgradnji optimalne metode potpornih vektora i ukoliko ih nema blizu hiper-ravni, klasifikacija će biti relativno laka. Ukoliko su podaci linearno nerazdvojivi koristi se nelinearni ***SVM***. Osnovna ideja jeste da se omogući da se podaci linearno razdvoje, a to se postiže preslikavanjem ulaznog vektorskog prostora u višedimenzionalni prostor. Razumijevanje načina rada ***SVM*** i ***k-NN*** algoritama izvršeno je na osnovu rada [6].

4. PREGLED IMPLEMENTACIJE I ANALIZA REZULTATA

Nakon detaljne pretrage više izvora podataka, pronađena su dva skupa podatka, koja predstavljaju dobru osnovu za dalju realizacija projekta. Prvi skup podataka sadrži osnovne podatke o aplikacijama, a obilježja u okviru njega su: ***Application name***, ***Category***, ***Rating***, ***Reviews***, ***Size***, ***Installs***, ***Type***, ***Price***, ***Content Rating***, ***Genres***, ***Last Updated***, ***Current Ver*** i ***Android Ver.*** Drugi skup podataka sadrži podatke o komentarima koje su korisnici mobilnih aplikacija ostavljali, a obilježja koja se mogu naći u okviru tog skupa su: ***App***, ***Translated_Review***, ***Sentiment***, ***Sentiment_Polarity***, ***Sentiment_Subjectivity***. Nakon pronalaženja odgovarajućih skupova podataka bilo je potrebno analizirati i razumjeti same podatke. Bilo je potrebno izbaciti nepotrebne podatke iz svakog pojedinačnog skupa, odrediti koja obilježja su bitna za predikciju i da li među njima ima koreliranih obilježja.

U prvom skupu podataka, koji sadrži osnovne podatke o aplikacijama, postojale su aplikacije koje nisu imale jedinstvene vrijednosti, postojali su duplikati aplikacija po imenu, pa su ti duplikati uklonjeni. Nad drugim skupom podataka, koji sadrži podatke o komentarima koje su korisnici mobilnih aplikacija ostavili, kao i o dodatnim informacijama u vezi značenja komentara primjenjen je isti postupak uklanjanja duplikata. Pod duplikatima su se smatrali aplikacije koje su imale u potpunosti iste vrijednosti obilježja u okviru drugog skupa podataka. Uklonjene su aplikacije koje imaju isto ime, komentar, sentiment, polaritet i subjektivnost. Nakon uklanjanja duplikata formiran je skup podataka spajanjem prethodno uređena dva skupa. Istraživanjem mašinskih algoritama, primjećeno je da u većini slučajeva isti bolje rade sa numeričkim vrijednostima tako da je svako obilježje nenumeričkog tipa konvertovano u numerički tip. Ciljno obilježje broj korisničkih preuzimanja/instaliranja mobilne aplikacije (***Installs***) je sadržalo karakter **+**, koje je uklonjeno i na taj način je konvertovan u brojnu vrijednosti (npr. 1000+ je konvertovano u 1000). Obilježje ***Price*** sadržalo je karakter ***$*** koji je uklonjen kako bi se vrijednost obilježja mogla konvertovati u numerički tip. Analizom vrijednosti obilježja ***Reviews*** uočeno je da jedna od vrijednosti sadrži karakter ***M***, za koju je pretpostavljeno da predstavlja oznaku za broj milion. Nakon uklanjanja karaktera, vrijednost obilježja je pomnožena sa milion kako bi se ostvarila konzistentnost sa ostalim vrijednostima u okviru ***Reviews*** obilježja. Radi očuvanja konzistentnosti, obilježje ***Size*** koje je bilo izraženo u ***MB*** (***megabyte***) i ***kB*** (***kilobyte***) je konvertovano u ***B*** (***byte***). Tip aplikacije je dobio numeričku vrijednost u zavisnosti od toga da li je u pitanju plaćena ili besplatna aplikacija. Za plaćenu aplikaciju vrijednost je 1, dok za besplatnu vrijednost je 0. Napravljene su izmjene u nazivima žanrova aplikacije (***Genres***) da se ne bi dva ista žanra tretirala kao dva različita (npr. ***Education***, ***Education*** se tretira kao ***Education***) i svakom jedinstvenom žanru se dodjeljuje odgovarajuća numerička vrijednost. Sličan postupak je sproveden nad kategorijom aplikacije (***Category***), obilježjem koje daje informaciju kojoj ciljnoj grupi korisnika je namijenjena aplikacija (***Content Rating***), kao i za ime aplikacije (***Name***). ***Sentiment*** komentara sadržao je nenumeričke vrijednosti ***Positive***, ***Neutral***, ***Negative***. Sve vrijednosti koje su imale pozitivan sentiment dobile su vrijednost 1. Vrijednosti sa neutralnim sentimentom zamijenjene su brojem 0, a za vrijednosti negativnog sentimenta dodijeljena je vrijednost -1.

Obilježja koja su uklonjena iz skupa podataka su datum posljednjeg ažuriranja (***Last Updated***), trenutna verzija aplikacije (***Current Ver***) i ***Android*** verzija (***Android Ver***), jer se na osnovu analize domena problema i rada [7] došlo do zaključka da navedena obilježja nisu potrebna za predviđanje popularnosti prilikom primjene izabranih mašinskih algoritama. Obilježje koje je izostavljeno u ovom skupu podataka su korisnički komentari (***Translated_Review***), jer su na osnovu njih već definisana ostala obilježja u samom skupu podataka. Prilikom primjene algoritama i stategija nad skupom podataka nad kojim su izvršene gore navedene izmjene, dobila se velika tačnost i mala standardna greška procjene. Sve to je ukazivalo da rezultati nisu realistični i da postoji određena zavisnost među obilježjima čije postojanje je određeno koeficijentom korelacije. Prilikom poređenja ciljnog obilježja ***Installs*** sa drugim obilježjima iz skupa podataka, najveći koeficijent korelacije dobijen je za obilježja ***Reviews*** i ***Size***, pa je zbog toga odlučeno da se navedena obilježja uklone iz skupa podataka. Na ovaj način je formiran novi skup podataka nad kojim su dalje primjenjeni odgovarajući algoritmi mašinskog učenja i strategije za rad sa nedostajućim vrijednostima.

Prilikom primjene strategija nad formiranim skupom podataka, kreiran je novi skup podataka za svaku primjenjenu strategiju. Zatim je svaki skup podataka podijeljen na trening i test skup u odnosu 70:30%. Nad trening skupom su primjenjeni klasifikatori ***k*** najbližih komšija – ***k-NN*** i mašine potpornih vektora (***Support Vector Machine – SVM***).

Koeficijent determinacije R^2 je korišćen kao mjera tačnosti koja pokazuje koliko promjene jedne promjenljive utiču na promjene druge promjenljive. Za evaluaciju rješenja korišćena je standardna greška procjene (***root mean squared error***). Što su obilježja više nezavisna, to je i manja standardna greška procjene.

4.1 Rezultati

U ovom poglavlju će biti prikazani rezultati koji su dobijeni primjenom različitih kombinacija strategija i algoritama u okviru Tabele 1.

Tabela 1. *Pregled rezultata*

Strategija	k-NN		SVM	
	Tačnost	Standardna greška procjene	Tačnost	Standardna greška procjene
ListWise	0.7418884433102443	71615195.2865384	0.3106088224571637	157315693.23087773
Dropping column	0.48601864181091875	140923004.58218968	0.3111003510470887	162417004.70113793
Mean	0.7415567122624379	103308206.14519987	0.3011741919864423	159221101.60640624
Median	0.743251422345963	103296162.41850609	0.30407940927248517	152529390.80998662
Mode	0.743251422345963	103300617.96097405	**0.29947948190291734**	150434220.04650003
Imputing with Constant	**0.7447040309889844**	103300611.1564436	0.3136424161723762	150510067.28282884
LOCF	0.7427672194649558	103308065.35110208	0.3165476334584191	157414701.41565114
NOCB	0.743251422345963	103314107.51389693	0.31013194528507443	154134711.7035762
Linearna interpolacija	0.7425251180244522	103323591.34203984	0.31678973489892265	165293235.85005927
Imputation using k-NN	0.7415567122624379	103308206.14519987	0.3153371262559012	155146761.0632167
MICE	0.7415567122624379	103308206.14519987	0.30492676431424764	149713477.606884

SVM algoritam je dao znatno lošije rezultate u odnosu na ***k-NN*** u kombinaciji sa bilo kojom strategijom. Najveća tačnost prilikom upotrebe ***SVM*** algoritma je 0.31678973489892265, a ***k-NN*** je pokazao bolji rezultat čak iako se izvrši poređenje sa najmanjom tačnosti upotrebom ***k-NN*** algoritma koja iznosi 0.48601864181091875. Najmanja tačnost dobila se prilikom upotrebe ***mode*** strategije i ***SVM*** algoritma (0.29947948190291734). Najveći koeficijent determinacije dobio se prilikom primjene ***k-NN*** algortima za klasifikaciju i imputacije sa konstantnim vrijednostima kao strategije za rad sa nedostajućim vrijednostima i on iznosi 0.7447040309889844.

5. ZAKLJUČAK

U radu je jasno definisan način na koji strategije za rad sa nedostajućim vrijednostima utiču na rješavanje konkretnog problema. Izvršena je podjela nedostajućih vrijednosti na određene tipove, kao i kratak pregled svakog od njih. Izvršen je kratak pregled metodologije koja je korišćena za pripremanje skupa podataka za sam proces predikcije. Predstavljena je podjela strategija za rad sa nedostajućim vrijednostima i opisivanje svake od njih. Osnovne osobine algoritama mašinskog učenja su, takođe, navedeni.

Izvršeno je izbacivanje duplikata, nepotrebnih i koreliranih obilježja, i određeno je koja obilježja su bitna za dalju analizu i upotrebu. Nad spojenim skupom podataka urađena je konverzija podataka u numeričke vrijednosti. Nad uređenim skupom podataka izvršena je primjena svih strategija koja je dovela do formiranja novih skupova podataka gdje se svaki od njih vezuje samo za jednu strategiju. Nad svakim skupom podataka dalje su se primjenjivali algoritmi mašinskog učenja namjenjeni za klasifikaciju.

SVM algoritam je dao znatno lošije rezultate u odnosu na ***k-NN*** u kombinaciji sa bilo kojom strategijom. Najmanji koeficijent determinacije se dobio prilikom upotrebe ***mode*** strategije i ***SVM*** algoritma, a najveća tačnost se dobila prilikom primjene ***k-NN*** algortima za klasifikaciju i linearne interpolacije kao strategije za rad sa nedostajućim vrijednostima.

6. LITERATURA

[1] D. B. Rubin, “Inference and missing data”, Biometrika 1976.

[2] https://www.kaggle.com/parulpandey/a-guide-to-handling-missing-values-in-python (pristupljeno u septembru 2020.)

[3] Z. Zhang, “Missing data imputation: focusing on single imputation”, 2016.

[4] H. Kang, “The prevention and handling of the missing data”, 2013.

[5] J. Zhang, D. Chen, “Interpolation calculation made EZ”

[6] Jasmina Đ. Novaković, “Rešavanje klasifikacionih problema mašinskog učenja”, 2013.

[7] G. Lee, T. S. Raghu; “Determinants of Mobile Apps Success: Evidence from App Store”, 2014.

Kratka biografija:

Sreten Petrović rođen je u Bijeljini 1995. god. Završio je gimnaziju JU Srednjoškolski centar “Vuk Karadžić” u Loparama 2014. godine. Osnovne akademske studije na Fakultetu tehničkih nauka u Novom Sadu završio je 2018. godine. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike odbranio je 2020.god.
Kontakt: sreten.95@hotmail.com