UNIVERZITET U BEOGRADU

RUDARSKO- GEOLOŠKI FAKULTET

Filip D. Arnaut

PRIMENA MODELA VOĐENIH PODACIMA U GEOFIZICI

Doktorska disertacija

Beograd, 2025

UNIVERSITY OF BELGRADE

FACULTY OF MINING AND GEOLOGY

Filip D. Arnaut

APPLICATION OF DATA- DRIVEN MODELS IN GEOPHYSICS

Doctoral Dissertation

Belgrade, 2025

Mentor:

dr Vesna Cvetkov, redovni profesor

Univerzitet u Beogradu, Rudarsko- geološki fakultet

Članovi komisije:

dr Dragana Đurić, vanredni profesor

Univerzitet u Beogradu, Rudarsko- geološki fakultet

dr Dragan Stankov, vanredni profesor

Univerzitet u Beogradu, Rudarsko- geološki fakultet

dr Aleksandra Kolarski, naučni saradnik

Univerzitet u Beogradu, Institut za fiziku u Beogradu

Datum odbrane: _____

ZAHVALNICA

Prikazana tema doktorske disertacije pod nazivom „*Primena modela vođenih podacima u geofizici*“ predstavlja višegodišnji rad i rezultat moje saradnje sa profesorima sa Katedre za geofiziku Rudarsko- geološkog fakulteta Univerziteta u Beogradu i istraživača iz Laboratorije za astrofiziku i fiziku jonosfere, Instituta za fiziku u Beogradu, Univerziteta u Beogradu u pronalaženju i istraživanju načina i primene modela vođenih podacima u geofizici.

Doktorska disertacija obuhvata više oblasti, te se autor posebno zahvaljuje:

- mentorki prof. dr Vesni Cvetkov na razumevanju, korisnim savetima i mentorstvu tokom četiri godine trajanja doktorskih studija i realizacije disertacije,
- dr Aleksandri Kolarski, na korisnim savetima i sugestijama tokom izrade doktorske disertacije i rukovođenjem vezano za poglavlje koje se odnosi na fiziku jonosfere Zemlje,
- prof. dr Dragani Đurić, na korisnim sugestijama, preporukama i savetima tokom doktorskih studija i izrade doktorske disertacije,
- Projektu Fonda za nauku Republike Srbije „*Karakterizacija i tehnološki postupci za reciklažu i ponovnu upotrebu flotacione jalovine rudnika Rudnik*“, čija su sredstva omogućila istraživanje jalovišta rudnika „Rudnik“ (projekat PRIZMA, broj 7522),
- Društvu istraživačkih geofizičara (eng. *Society of Exploration Geophysicists*- SEG) na finansijskoj pomoći u vidu stipendija 2022. godine (eng. *Normal and Shirley Domenico Scholarship*) i 2023. godine (eng. *Chevron Scholarship*),
- članovima komisije na korisnim savetima i sugestijama tokom realizacije disertacije, koji su doprineli kvalitetu rada.

Takođe, zahvaljujem se svim mrežama otvorenih podataka koji su primenjeni u ovoj disertaciji- *Worlwide Archive of Low- Frequency Data and Observations* (WALDO), Agenciji za zaštitu životne sredine (SEPA), *OpenStreetMaps*, Iowa State Mesonet (Iowa State University) i drugima.

Izražavam posebnu zahvalnost Simoni Jevremović, Vladanki i Dušanu Arnaut za neizmernu pomoć i podršku tokom izrade doktorske disertacije. Njihova podrška i ohrabrenje bili su oslonac bez kojeg realizacija disertacije ne bi bila moguća.

PRIMENA MODELA VOĐENIH PODACIMA U GEOFIZICI

REZIME

Modeli vođeni podacima karakterišu se ubrzanim razvojem zbog široke pristupačnosti i unapređenja sposobnosti procesora, softverskih paketa i računarske memorije. Primena metoda mašinskog učenja, prognoziranja i analize vremenskih serija aktivno napreduje unutar šire naučne oblasti nauke o podacima. Potencijalna područja primene ovih metoda na geofizičke, geološke i podatke koji se odnose na blisku Zemljinu okolinu su brojna i raznovrsna. Doktorska disertacija prikazuje kombinovane rezultate različitih primena prethodno pomenutih metoda, i to: primena metoda prognoziranja vremenskih serija na podatke koncentracije zagađujućih materija u vazduhu (Fejsbuk Profet model), primena metoda mašinskog učenja za imputaciju podataka koncentracije zagađujućih materija u vazduhu (dvosmerna imputacija podataka), primena metoda klasifikacije mašinskog učenja za prostornu klasifikaciju ofiolita istočne Vardarske zone (Severna Makedonija), primena metoda mašinskog učenja na podatke signala vrlo niskih frekvencija koji se prostiru subjonosferski za detekciju poremećaja amplitude pomenutih signala, primena metoda mašinskog učenja za dobijanje talasovodnih parametara oblasti D Zemljine jonosfere tokom poremećenih jonosferskih uslova usled solarnih flerova i primena statističkih metoda na podatke magnetne susceptibilnosti dobijene od uzoraka materijala sa flotacijskog jalovišta rudnika „Rudnik“ (Srbija). Svaki od prethodnih primera prikazuje primenu različitih metoda nauke o podacima (mašinsko učenje i metode prognoziranja vremenskih serija) i primenjene statistike na realne podatke, kako bi se dobile informacije koje nije moguće dobiti konvencionalnim metodama. Disertacija takođe prikazuje značaj budućih unapređenja i primene ovih metoda na geo-podatke, koji mogu imati veliki značaj za istraživače i industriju u datim domenima.

Ključne reči: mašinsko učenje, prognoziranje vremenskih serija, prostorno prognoziranje, nauka o podacima, analiza (geo)podataka.
Naučna oblast: Geo-nauke
Uža naučna oblast: Geofizika
UDK:
621:004.85(043.3)
001.103

APPLICATION OF DATA-DRIVEN MODELS IN GEOPHYSICS

ABSTRACT

Data-driven models have rapidly expanded due to widespread accessibility and advancements in processing capabilities, software packages, and computer memory. The utilization of machine learning and time-series forecasting and analysis techniques is actively progressing within the broader scientific domain of data science. The potential application of these methods to geophysical, geological, and near-Earth physics datasets is extensive and diverse. This PhD dissertation presents the combined results from various applications, specifically: the utilization of time-series forecasting techniques (Facebook's Prophet algorithm) on particulate matter concentration data; the employment of machine learning methods to impute missing observations (bi-directional data imputation) for particulate matter concentration data; the application of machine learning classification techniques for the spatial classification of ophiolites in the East Vardar Ophiolite Zone, North Macedonia; the implementation of machine learning methods on subionospheric very low frequency signal data for amplitude anomaly detection; the use of machine learning methods to derive D region ionospheric waveguide parameters under disturbed ionospheric conditions due to solar flares; and the application of statistical methods on magnetic susceptibility data acquired from a flotation mine tailing material from the mine "Rudnik" (Serbia). Each of the aforementioned examples illustrates the application of diverse data science (machine learning and time- series forecasting methods) and applied statistical methodologies on real-world data to extract insights that are not conventionally attainable. The dissertation emphasizes the significance of future advancements and applications of these methods for geo-data, as they are highly beneficial for researchers and industry professionals in these domains.

Keywords: machine learning, time- series forecasting, spatial forecasting, data science, (geo)data analysis.
Scientific field: Geosciences
Scientific subfield: Geophysics
UDC:
621:004.85(043.3)
001.103

Sadržaj

1. Uvod

Geofizika proučava fizičke procese, pojave i fizička svojstva Zemlje i njenog okolnog prostora. U tu svrhu se koristi skup metoda zasnovanih na analizi i tumačenju fizičkih parametara i svojstava, koji su povezani sa njihovim geološkim, petrološkim i drugim karakteristikama, značajnim za cilj istraživanja. Dobijeni (geofizički) podaci mogu biti prostorni, vremenski ili prostorno- vremenski, u zavisnosti od potreba i primenjene metode istraživanja, i u odnosu na način prikupljanja podataka.

Primena modela vođenih podacima (eng. *Data driven models*) predstavlja proces modelovanja određenog sistema sa izmerenim, realnim podacima, za razliku od modelovanja zasnovanog na fizici same pojave i modelovanju kroz matematičke jednačine (eng. *Physics based models*). U osnovi modela vođenih podacima je analiza svih dostupnih podataka koji su vezani za predmet istraživanja, sa posebnom pažnjom posvećenom pronalaženju veza između raznorodnih podataka kod kojih uzajamna interakcija nije eksplicitna ili poznata. Ono što ograničava ovaj vid modelovanja je kvalitet i kvantitet korišćenih podataka. Modelovanje vođenim podacima u osnovi obuhvata mašinsko učenje, prognoziranje i analizu vremenskih serija, ali i brojne druge statističke metode, a zajedno sa metodama pripreme, skladištenja i transformacije podataka, ima veliki potencijal primene na geofizičke i druge podatke vezane za Zemljino okruženje.

1.1. Cilj disertacije

Cilj disertacije je primena izabranih metoda mašinskog učenja, prognoziranja i analize vremenskih serija i statističkih metoda na različite prostorne i vremenske geofizičke podatke. Ovakav pristup omogućava prikupljanje dodatnih informacija iz podataka koji nisu dostupni standardnim metodama istraživanja, kao i automatizaciju određenih procesa koji zahtevaju veliki trud i uloženo vreme istraživača. U tu svrhu izdvojeno je pet predmeta istraživanja, i to

a) Primena metoda prognoziranja vremenskih serija i metoda mašinskog učenja na podatke koncetracije zagađujuće materije u vazduhu u Republici Srbiji (Arnaut et al., 2023a; 2024a);

b) Primena klasifikacionih metoda mašinskog učenja za prognoziranje prostornog položaja ofiolita istočne Vardarske zone na prostoru Severne Makedonije (Arnaut et al., 2024b; Arnaut, 2024);
c) Primena klasifikacionih metoda mašinskog učenja za klasifikaciju poremećenog signala vrlo niskih frekvencija koji se prostire sub-jonosferski (eng. *Very Low Frequency, VLF*) (Arnaut et al., 2023b, 2024c, 2025);
d) Primena regresionih metoda mašinskog učenja za određivanje parametara niske jonosfere odnosno talasovodnih parametara parametara oblasti D jonosfere (Arnaut et al., 2023c);
e) Primena statističkih metoda na podatke magnetne susceptibilnosti prikupljene iz tri bušotine na flotacijskom jalovištu rudnika „Rudnik“.

1.2. Polazne hipoteze

Na osnovu prethodno definisanog predmeta i ciljeva doktorske disertacije postavljene su sledeće polazne hipoteze:

a) Prognoziranje vremenskih serija koncentracije zagađujućih materija u vazduhu ima značajnu primenu u dobijanju budućih vrednosti odabranih parametara na određenoj lokaciji. Ova informacija je od posebne važnosti za osobe sa respiratornim oboljenjima, kao i za najmlađu i najstariju populaciju koja je najviše ugrožena visokim nivoima zagađenja vazduha. Iako trenutno ne postoji stabilna, univerzalna i u potpunosti tačna metoda za prognozu koncentracije zagađujućih materija u vazduhu, u disertaciji je prikazana primena Fejsbuk Profet (eng. *Facebook Prophet*) algoritma za prognoziranje vremenskih serija na podacima o koncentraciji zagađujućih materija u vazduhu sa merne stanice Beograd- Zeleno brdo. Iako je ovaj algoritam prvobitno razvijen za potrebe prognoze upotrebe društvenih mreža, našao je široku primenu kako u industriji tako i u brojnim naučno- istraživačkim granama. U okviru ove disertacije biće testirana uspešnost pomenutog algoritma za prognoziranje koncentracije zagađujućih materija u vazduhu, što spada u naučnu oblast fizike atmosfere. Takođe, metode mašinskog učenja primenjene su i za aproksimaciju preskočenih opservacija u podacima o koncentraciji zagađujućih materija u vazduhu. Sa obzirom na to da su podaci o koncentraciji zagađujućih materija prikupljeni kontinuiranim monitoringom, moguće je da tehnički problemi izazovu preskočene opservacije. Razvijanje novih metoda za rešavanje takvih

nedostataka u podacima, dovodi do značajnog napretka u tačnosti, pouzdanosti i sveobuhvatnom kvalitetu analiziranih podataka.

b) Kartiranje litoloških jedinica je često vremenski zahtevan proces koji uključuje širok spektar terenskih aktivnosti i obimnu obradu podataka. Korišćenje metoda mašinskog učenja može značajno unaprediti ovaj proces, jer omogućava modelima da "nauče" prostorni raspored litoloških jedinica na temelju podataka prikupljenih iz jednog dela istraživanog područja, a zatim da primene ovu logiku na neistraženi deo područja. U disertaciji biće primenjeni klasifikacioni modeli mašinskog učenja, kao što su model slučajnih šuma (eng. *Random Forest*), model K- najbližih suseda (eng. *K-Nearest Neighbours*) i model ekstremnog povećanja gradijenta (eng. *Extreme gradient boosting*), kako bi se prognozirala prostorna raspodela ofiolita u istočnoj Vardarskoj zoni (Severna Makedonija). Polazna hipoteza istraživanja odnosi se na analizu odnosa između binarne ciljne promenljive (ofioliti i ne- ofioliti) i njenog uticaja na performanse modela. U tom kontekstu, istraživaće se da li modeli koji koriste manje podataka za učenje, ali sa uravnoteženijim odnosom klasa ciljne promenjive, mogu postići bolje rezultate u klasifikaciji geo- prostornih podataka. Ova istraživačka postavka pruža mogućnost da se testira efikasnost modela u različitim odnosima raspodele klasa ciljne promenljive, što može značajno doprineti tačnosti klasifikacija vezano za geološka istraživanja i kartiranje. Takođe, primena kombinacije geofizičkih (dubinski vezanih) i satelitskih (površinski vezanih) podatka predstavlja novitet u oblasti primene metoda mašinskog učenja za klasifikaciju litologije.

c) Registracija podataka niske jonosfere upotrebom radio signala vrlo niskih frekvencija (eng. *Very Low Freqnecy*- VLF signali, 3-30 kHz) se vrši u visokoj minutnoj rezoluciji. Pored tehničkih smetnji i smetnji koje su svojstvene za većinu signala merenih u prirodi, VLF signal je dodatno opterećen uticajem efekata solarnih flerova, kao jednim od najznačajnijih uzročnika poremećaja jonosfere Zemlje. Količina podataka koju istraživač mora da obradi zavisi od broja analiziranih parova antena predajnik-prijemnik i dužine ispitivanog perioda, i često može biti izrazito velika. Da bi se proces obrade podataka znatno ubrzao primeniće se metode klasifikacije mašinskog učenja na podatke VLF signala sa ciljem automatizacije procesa otklanjanja poremećenog dela signala. Pored pomenute automatizacije procesa obrade podataka, drugi cilj koji može proisteći iz budućih istraživanja vezano za ovu problematiku je razvijanje algoritma za prepoznavanje karakteristika signala u (skoro) realnom vremenu.

d) Najniža oblast jonosfere Zemlje, oblast D (50-90 km), u smislu istraživanja primenom tehnologije prostiranja VLF signala, određen je sa dva jonosferska parametra, oštrinom i visinom granice reflektovanja VLF signala. Tokom neporemećenih jonosferskih uslova vrednost oba pomenuta parametra je relativno konstantna. Pod dejstvom solarnih flerova, u poremećenim uslovima, dolazi do značajne promene vrednosti oštrine i visine granice reflektovanja VLF signala. Numeričkim modelovanjem koje se bazira na modelovanju zasnovanim na fizici same pojave mogu se odrediti vrednosti ovih parametara za poremećeni period, međutim, zbog kompleksnosti samog procesa to je izazovan zadatak i dugotrajan proces. Iz tog razloga su testirane regresione metode mašinskog učenja na podatke oblasti D jonosfere i mogućnost određivanja oštrine i visine granice reflektovanja VLF signala tokom poremećenih jonosferskih uslova pod uticajem solarnih X- flerova.
e) Merenja magnetne suseptibilnosti visoke rezolucije po vertikalnom profilu (sa dubinom) tela flotacijskog jalovišta rudnika „Rudnik“ obezbediće veliki skup podataka na koji će se primeniti statističke metode u cilju međusobnog poređenja materijala jalovišta iz tri bušotine, a takođe i primenu metoda statistike za određivanje zona od značaja unutar bušotina. Laboratorijska merenja magnetne suseptibilnosti niskog polja na uzorkovanom materijalu iz tri bušotine dubine oko 15 m na svakih 10 cm omogućiće razgraničenje zona obogaćenja, kao i planiranje dodatnih mineraloških i geohemijskih istraživanja u zonama povećane magnetne suseptibilnosti, koje ukazuje na visoko obogaćenje teškim metalima.

1.3. Sažetak disertacije

Zbog široke oblasti primene i brojnosti metoda mašinskog učenja, metoda prognoziranja i analize vremenskih serija i pratećih metoda obrade, transformacije i statističkih metoda, disertacija je podeljena u sledeća poglavlja na osnovu oblasti primene pomenutih metoda:

a) **Poglavlje 1**- je uvod u kome su ukratko prikazane teme istraživanja, ciljevi i polazne hipoteze;
b) **Poglavlje 2-** u okviru ovog poglavlja koje se bavi metodologijom sprovedenog istraživanja prikazan je radni tok uopštene primene metoda mašinskog učenja (klasifikacije i regresije), primena metoda prognoziranja i analize vremenskih serija kao i različitih statističkih metoda. Poglavlje metodologije je prikazano u opštim uslovima

kako bi čitalac dobio jasnu sliku da su metode mašinskog učenja i druge primenjene metode usko povezane sa statistikom i sa namerom da se poglavlje metodologije prikaže na intuitivan način.

c) **Poglavlje 3-** treće poglavlje se bavi rezultatima sprovedenih istraživanja. Rezultati su grupisani prema oblastima istraživanja i prikazani u četiri potpoglavlja;
d) **Poglavlje 4-** diskusija rezultata i predlog budućih istraživanja prikazani su u četvrtom poglavlju i
e) **Poglavlje 5-** u kome su data zaključna razmatranja ove doktorske disertacije.

Disertacija predstavlja skup četiri različite grupe primene metoda vođenih podacima i rezultat je višegodišnjeg rada vezanog za pronalaženje i primenu ovih metoda na različite vrste (geo)podataka. Cilj disertacije je da prikaže obrađena područja primene pomenutih metoda. Stoga, prilikom prikaza primera obrađenih u disertaciji nije se ulazilo u detalje. Kako bi disertacija ostala fokusirana na glavnu temu, svako poglavlje sadrži kratak uvodni deo koji čitaocu pruža osnovne informacije o datoj pojavi, ali ne i detaljan pregled mehanizma te pojave, njenog nastanka ili drugih aspekata.

Prilikom pisanja disertacije, u velikoj meri, a takođe u određenim delovima u potpunosti, korišćeni su radovi: Arnaut et al., 2023a, b, c; Arnaut et al., 2024a, b, c, Arnaut et al., 2025 i Arnaut, 2024.

2. Teorijske osnove i metodologija

Modelovanje vođeno podacima (eng. *Data driven modeling*) predstavlja skup statističkih metoda koje se koriste za modelovanje odabranih procesa na osnovu izmerenih podataka i deo je šireg okvira nauke o podacima (eng. *Data science*). Metode u okviru modelovanja vođenog podacima pružaju širok spektar informacija, kako u oblasti prirodnih, tako i u oblasti društvenih nauka. Nagli razvoj računarskih tehnologija i povećana dostupnost velike količine računarske memorije doprinose značajnoj ekspanziji i rastu interesa za ove metode, koje privlače istraživače iz različitih oblasti nauke i privrede. Trenutno, metode modelovanja vođenog podacima primenjuju se u gotovo svim oblastima nauke i privrede. Neki od poznatih primera uključuju: klasifikaciju neželjenih elektronskih poruka (eng. *Spam emails*) (Dada et al., 2019), velike jezičke modele (eng. *Large language models*) (Naveed et al., 2023), prognozu potrošačke aktivnosti (Žunić et al., 2020), klasifikaciju različitih vrsta malignih bolesti (ZainEldin et al., 2022; Zhang et al., 2023), automobile koji se samostalno voze (Chougle et al., 2023), i druge. Zajednička osobina ovih i sličnih primera je upotreba postojećih podataka za modelovanje i dobijanje dodatnih informacija o određenom procesu ili automatizaciji datog procesa.

Pored metoda zasnovanih na izmerenim podacima, postoji i drugi tip modela, tzv. numerički modeli, koji spadaju u modele zasnovane na fizici (eng. *Physics based modeling*). Za ovaj tip modela izmereni podaci nisu neophodni, jer se koriste matematičke jednačine za modelovanje željenih procesa. Zbog toga numerički modeli mogu imati ograničenu primenu u rešavanju određenih problema koji zahtevaju rad sa realnim podacima. Važno je napomenuti da modeli vođeni podacima takođe imaju ograničenu primenu, koja najviše zavisi od količine i kvaliteta podataka. Ukoliko nije moguće dobiti izmerene podatke ili je kvalitet izmerenih podataka upitan, rezultati modelovanja će biti upitnog kvaliteta. Numeričko modelovanje i modelovanje vođeno podacima imaju svoja specifična područja primene, koja zavise od vrste informacija koje je potrebno dobiti, kao i od vrste, tipa i kvaliteta korišćenih podataka. Izbor između ova dva pristupa zavisi od dostupnih podataka, vrste i cilja istraživanja.

Takođe, potrebno je prikazati i prelazne slučajeve kada modelovanje koristi numeričke metode, ali spada pod modelovanje koje je vođeno realnim, odnosno izmerenim podacima. Takav slučaj je u geofizičkoj inverziji, a najpre se odnosi na geofizičku elektrometrijsku inverziju. Prilikom

elektrometrijske inverzije, izmereni podaci predstavljeni su na sekciji izmerenih prividnih specifičnih električnih otpornosti, dok je cilj numeričkog modelovanja da se dobije direktan model koji proizvodi sekciju teoretski sračunatih prividnih specifičnih električnih otpornosti, koja najmanje odstupa od sekcije izmerenih prividnih specifičnih električnih otpornosti. Numeričko modelovanje u ovom slučaju koristi se radi optimizacije dobijanja modela nakon inverzije, ali je proces po svojoj prirodi vođen podacima. Sa druge strane, prilikom direktnog modelovanja, gde se zadaje model podpovršine, a očekuje se sekcija teoretski sračunatih prividnih specifičnih električnih otpornosti, postupak nije vođen podacima, već spada pod modele zasnovane na fizici procesa. Prema tome, moguće je svrstati geofizičko direktno modelovanje pod modelovanje zasnovano na fizici, dok geofizičko inverzno modelovanje spada pod modelovanje vođeno podacima, prema načinu i prirodi podataka koji se koriste i generišu.

Metode veštačke inteligencije (eng. *Artificial Intelligence*), uključujući metode mašinskog učenja (eng. *Machine Learning*) i metode dubokog učenja (eng. *Deep Learning*), trenutno su u ekspanziji i predstavljaju vrlo aktuelne metode modelovanja vođenog podacima u različitim oblastima nauke i privrede. Ove metode su u velikom razvoju i privlače značajno interesovanje zbog širenja i rasta kompjuterskih kapaciteta. Međutim, važno je napomenuti da ove metode nisu nove, jer one postoje još od polovine XX veka, ali su postale dostupnije širem spektru istraživača i drugih korisnika zahvaljujući povećanim računarskim kapacitetima. Oblasti primene metoda veštačke inteligencije, mašinskog učenja i sličnih metoda su vrlo široke, od prethodno pomenute klasifikacije neželjene elektronske pošte, do različitih modelovanja u medicinske svrhe i drugih primera koji su već navedeni.

Pod modelovanjem vođenim podacima potpadaju i modeli prognoziranja vremenskih serija (eng. *Time-series forecasting*). Prognoziranje vremenskih serija predstavlja skup različitih metoda koje koriste izmerene podatke u vremenu za dobijanje budućih (trenutno neizmerenih) podataka. Metode prognoziranja vremenskih serija razvijaju se od početka XX veka (Bisgaard & Kulachi, 2011) i našle su primenu u različitim oblastima nauke i privrede, kao što su: prognoziranje prenosa virusa COVID-19 (Papastefanopoulos et al., 2020; Somyanonthanakul et al., 2023), prognoziranje u oblasti energetike (Shakeel et al., 2023a, b), prognoziranje stepena inflacije (Meyler et al., 1998), i dr. Analiza i prognoziranje vremenskih serija našli su široku primenu u ekonomiji, te je često slučaj da se prognoziranje vremenskih serija izučava i razvija u okviru polja ekonometrije, discipline koja se bavi primenom statističkih metoda

(primenjene statistike) na ekonomske podatke. Metode prognoziranja i analize vremenskih serija predstavljaju značajne statističke metode koje se sve više primenjuju i u drugim naučnim disciplinama i oblastima.

Metode modelovanja vođenim podacima našle su primenu u oblastima geonauka, gde su metode mašinskog učenja korišćene za klasifikaciju različitih litologija i mineralnih sirovina (Zuo & Carranza, 2023; Carranza & Laborte, 2015a, b; Cracknell & Reading, 2014), kao i za klasifikaciju različitih vrsta vegetativnog sloja ili prekrivača zemljišta (Huang et al., 2002; Foody & Mathur, 2004; Ham et al., 2005; Waske & Braun, 2009; Cracknell & Reading, 2014). Područja primene metoda vođenih podacima u geonaukama su širokog opsega i tema su ovog rada, pa će detaljan prikaz biti posvećen u narednim poglavljima.

Poglavlje metodologije daje detaljan prikaz metoda mašinskog učenja- klasifikacija, regresija, transformacije i obrada podataka, modeli, kvantifikacija odstupanja modela i dr., kao i metoda prognoziranja vremenskih serija- transformacije i obrada podataka, pretpostavke različitih modela i druge specifičnosti koje su vezane za modelovanje i prognoziranje vremenskih serija.

Na kraju, potrebno je naglasiti da je većina literature koja se bavi mašinskim učenjem i prognoziranjem vremenskih serija, zasnovana na matematičkoj notaciji, prikazujući kako različiti modeli funkcionišu kroz ovu notaciju. Iako to nije pogrešan pristup, on nije najpraktičniji za različite grane koje primenjuju ove modele, a ne bave se razvojem samih modela. Zbog toga će prikaz modela u ovoj disertaciji biti prikazan kroz uopšteni radni tok modelovanja. Time će se omogućiti bolje razumevanje načina funkcionisanja modela i generalno mašinskog učenja, bez preteranog zalaženja u matematičku notaciju. Ovim pristupom smatra se da će čitaocu biti pruženo intuitivno razumevanje procesa modelovanja, kao i razlike između različitih modela. Na ovaj način dat je uvid u svaki primenjeni postupak, sa nastojanjem da se ogradi od negativne konotacije vezano za mišljenje da je mašinsko učenje "crna kutija" (eng. *Black box*), jer u osnovi predstavlja skup relativno jednostavnih i intuitivnih statističkih postupaka koji su doprineli velikom razvoju u različitim granama nauke i privrede.

2.1. Metode nadgledanog mašinskog učenja

Kao što je prethodno napomenuto, metode mašinskog učenja nisu nove; one su poznate još od četrdesetih godina XX veka, u radovima Alana Turinga (eng. *Alan Turing*). Već pedesetih godina XX veka razvijen je prvi model- Perseptron (eng. *Perceptron*), koji se smatra pretečom današnjih neuronskih mreža (Nikolić & Zečević, 2019). Razvojem kompjuterskih tehnologija, kao što su brži i moćniji procesori, veće količine računarske memorije i drugi faktori, mašinsko učenje je početkom dvehiljaditih godina doživelo nagli razvoj. Trenutno, mašinsko učenje ima široku primenu u različitim oblastima, kako u nauci, tako i u privredi.

2.1.1. Osnovni pojmovi

Mašinsko učenje u svojoj osnovi predstavlja skup različitih metoda, tehnika, algoritama i procedura koje imaju cilj da „nauče“ mašinu na određenom skupu podataka, kako bi nakon toga mogla da primeni stečeno „znanje“ na drugom skupu podataka, bez direktne komunikacije o međusobnoj povezanosti između podataka ili unutar samih podataka. Budući da mašinsko učenje nije predstavljeno na najbolji mogući način, pre svega zbog svoje sveobuhvatnosti, preciznije definicije biće date u daljem tekstu. Dobar primer cilja mašinskog učenja navode Nikolić i Žečević (2019), gde se kao primer koristi jednačina $F = m \times a$, koja predstavlja povezanost sile F (N), mase m (kg) i ubrzanja a (ms^{-1}). Do ove jednačine se došlo empirijskim putem, tj. čovek je morao da prepozna povezanost između tri parametra i dodeli odgovarajuću matematičku notaciju. Mašinsko učenje, sa druge strane, ima drugačiji cilj. Na primer, ako se ne zna povezanost između sile, mase i ubrzanja, moguće je odrediti ciljnu promenljivu (eng. *Target variable*)- u ovom slučaju silu- i atribute (eng. *Features*)- masa i ubrzanje. Uz odgovarajuće podatke i primenu mašinskog učenja, moglo bi se doći do saznanja o povezanosti atributa sa ciljnom promenljivom, bez prethodnog znanja o vrsti povezanosti.

Ciljna promenljiva u sklopu mašinskog učenja predstavlja vrednost koja se modeluje, odnosno izlaz modela mašinskog učenja. Definisanje ciljne promenljive jedan je od prvih koraka u modelovanju mašinskim učenjem, gde istraživač određuje koji je parametar od značaja koji treba modelovati, u zavisnosti od problema koji treba rešiti. U većini slučajeva ciljna promenljiva je predstavljena jednim parametrom, ali može biti i kombinacija dva ili više parametara. Naredni korak u modelovanju mašinskim učenjem je određivanje atributa, odnosno grupa podataka koji opisuju ciljnu promenljivu. Ovaj korak se često naziva otkrivanje

atributa (eng. *Feature discovery*) ili inženjering atributa (eng. *Feature engineering*). Koraku otkrivanja atributa biće posvećeno posebno pažnje u narednim poglavljima, jer se smatra da je taj korak u trenutno dostupnoj literaturi nedovoljno razrađen i delimično zapostavljen, a predstavlja jedan od ključnih koraka prilikom modelovanja mašinskim učenjem.

U prethodnom prikazu ciljne promenljive i atributa može se uvesti novi pojam koji se odnosi na vrstu modela mašinskog učenja. Nadgledano (eng. *Supervised*) mašinsko učenje predstavlja vrstu mašinskog učenja u kojoj su u podacima poznati ulazi (atributi) i izlaz (ciljna promenljiva). Ova vrsta mašinskog učenja podrazumeva postojanje ciljne promenljive koja predstavlja izlaz za atribute. Drugim rečima, nadgledano mašinsko učenje mapira vrednosti atributa na izlaz, čime se modeluje izlazna promenljiva. Drugi tip mašinskog učenja je nenadgledano (eng. *Unsupervised*) mašinsko učenje, gde ne postoji ciljna promenljiva, već se od mašine očekuje da sama „donese zaključke“ o podacima. Najveću primenu je pronašlo nadgledano mašinsko učenje, a koje je i korišćeno u svim primerima ove doktorske disertacije.

Prema vrsti podataka, razlikuju se numerički podaci (eng. *Numerical data*)- celi brojevi (eng. *Integer*) i brojevi sa zarezom (eng. *Float*); kategorične podatke (eng. *Categorical data*)- nominalne (eng. *Nominal*) klase, koje nemaju određeni međusobni rang (npr. boje, imena itd.), i ordinalne (eng. *Ordinal*) klase, koje imaju određeni međusobni rang (npr. nisko, srednje, visoko, vrlo visoko itd.); te binarne (eng. *Binary*) podatke, koji su obično predstavljeni sa 1 ili 0, “da” ili “ne”, itd. U zavisnosti od vrste ciljne promenljive, modeli mašinskog učenja mogu se koristiti u svrhu regresije, ukoliko je ciljna promenljiva predstavljena numeričkim podacima, ili u svrhu klasifikacije, ukoliko je ciljna promenljiva predstavljena kategoričnim ili binarnim podacima. Različite svrhe modela (klasifikacija ili regresija) zahtevaju različite postupke transformacije, obrade i različite mere ocene kvaliteta modela.

2.1.2. Radni tok nadgledanog mašinskog učenja

Nakon definisanja osnovnih pojmova u mašinskom učenju i podela mašinskog učenja na klasifikaciju i regresiju, od kojih svaki ima svoje specifičnosti, moguće je prikazati generalizovani radni tok modelovanja mašinskim učenjem (Slika 1). Kroz dati radni tok biće prikazane različite vrste obrade, transformacije, mere ocene kvaliteta modela, kao i sami modeli mašinskog učenja. Određeni pojmovi koji su prikazani u okviru generalizovanog

radnog toga mašinskog učenja se takođe mogu videti u drugim izvorima (npr. Samardžić-Petrović, 2014; Brink et al., 2016).

2.1.2.1. Definisanje problema nadgledanog mašinskog učenja

Definisanje problema nadgledanog mašinskog učenja predstavlja jedan od prvih koraka u radnom toku mašinskog učenja. Prilikom ovog koraka, početni podaci se analiziraju kako bi se stekao uvid u vrstu mašinskog učenja koje će biti primenjeno- klasifikaciju ili regresiju. Takođe, u ovom koraku se analizira i količina podataka koja je dostupna istraživaču, te se odlučuje da li je potrebno proširiti set podataka i/ili povećati njihov kvantitet. Što se tiče kvantiteta podataka, ne postoji tačno definisan broj podataka koji predstavlja granicu za korišćenje metoda mašinskog učenja. Potreban kvantitet podataka u velikoj meri zavisi od problema koji se istražuje, kvaliteta podataka (gde je potreban veći broj podataka ako su podaci opterećeni šumom), odabrane metode (pri čemu metode dubokog učenja i neuronske mreže zahtevaju veći broj podataka nego klasične metode mašinskog učenja), kao i varijacija u samim podacima i međusobnog odnos ciljne promenljive i atributa. Takođe, može se voditi logikom "*što više podataka, to bolje*", što predstavlja validnu pretpostavku u okviru mašinskog učenja. Međutim, istraživač treba voditi računa o trenutku kada se dostigne tačka opadajućeg doprinosa (eng. *Point of diminishing returns*), gde dodavanje novih uzoraka ne povećava kvalitet modela, a ujedno povećava utrošeno računarsko vreme za modelovanje.

2.1.2.2. Otkrivanje atributa

Nakon odabira klasifikacije ili regresije, u zavisnosti od cilja istraživanja i ciljne promenljive, naredni korak je otkrivanje atributa. Otkrivanje atributa predstavlja jedan od najvažnijih koraka u mašinskom učenju, te će tom koraku kroz ceo rad biti posvećena posebna pažnja. Na samom početku, potrebno je detaljno definisati značaj atributa za mašinsko učenje. Prilikom modelovanja ciljne promenljive, modeli mašinskog učenja modeluju ciljnu promenljivu prema varijacijama atributa. Time se dolazi do zaključka da upravo atributi imaju ključnu ulogu u uspešnosti modelovanja ciljne promenljive. Zbog toga, odabir atributa treba da se vrši sa velikom pažnjom, uz adekvatan izbor atributa prema domenskom znanju o pojavi koja se modeluje. Kao primer za odabir atributa može se uzeti modelovanje koncentracije zagađujućih materija u vazduhu, na primer, koncentracija čestica do 2.5 $\mu m/m^3$ ($PM_{2.5}$). Prilikom modelovanja koncentracije finih čestica u vazduhu (ciljna promenljiva), potrebno je analizirati

koji drugi parametri mogu uticati na njeno povećanje ili smanjenje, odnosno varijaciju. Stoga, u literaturi o modelovanju ili prognoziranju koncentracije zagađujućih materija u vazduhu mašinskim učenjem često se susreću atributi poput: drugih parametara koncentracije zagađujućih materija u vazduhu (koncentracija čestica do 10 $\mu m/m^3$, koncentracija NO_2, SO_2, CO, O_3 i dr.), meteoroloških podataka (temperatura, vlažnost vazduha, padavine, pritisak i dr.), podataka o saobraćaju u blizini merne stanice (jer je poznato da intenzitet saobraćaja utiče na koncentraciju zagađujućih materija u vazduhu), kao i podataka o vremenu (vreme u danu, dan u nedelji, nedelja u godini, mesec u godini itd.), pošto je poznato da ljudska aktivnost zavisi od perioda u danu, nedelji i mesecu, što se takođe može odraziti na preciznost modela mašinskog učenja. Prethodni primer ilustruje koji svi parametri mogu poslužiti kao atributi za modelovanje mašinskim učenjem. Uopšteno, za druge probleme koji se mogu rešavati, potrebno je odrediti relevantne atribute za datu ciljnu promenljivu.

Nakon otkrivanja svih relevantnih atributa, a pre podele seta podataka na trening- test ili trening- test- validacija, moguće je, po potrebi, izvršiti progušćavanje podataka, odnosno povećanje kvantiteta podataka u setu. Ovaj korak treba preduzeti sa dozom rezerve i odlučiti se za njega samo ako nije moguće povećati broj uzoraka na neki drugi način. Pošto će u narednim poglavljima biti reči o povećanju broja uzoraka, ovde će biti prikazana jedna od mogućih metoda za takvu vrstu obrade podataka. Progušćavanje uzoraka moguće je uraditi na osnovu raspodele vrednosti za dati atribut i/ili ciljnu promenljivu. Jedna od metoda uključuje povećanje broja uzoraka primenom ocene gustine raspodele pomoću funkcije jezgra (eng. *Kernel density estimation*). Upotrebom ove metode ocenjuje se raspodela odabranog parametra, nakon čega se generišu sintetički podaci iz originalne raspodele, pri čemu broj sintetičkih podataka definiše istraživač. Drugim rečima, ocenjuje se raspodela odabranog parametra, a zatim se iz nje izvlači *n* sintetičkih uzoraka. Nakon dobijanja seta sintetičkih podataka, potrebno je izvršiti test poklapanja originalne raspodele i sintetičke raspodele. Za tu svrhu može se koristiti neparametarski Kolmogorov-Smirnov test (Berger & Zhou, 2014), koji proverava da li dva seta podataka prate istu raspodelu. Interpretacija Kolmogorov-Smirnov testa zasniva se na p-vrednosti (eng. *p-value*), pri čemu, ako je *p- vrednost* veća od 0,05, ne postoji dovoljno statističkog osnova da se odbije nulta hipoteza, što znači da su raspodele verovatno slične ili iste. Sa druge strane, ako je p-vrednost manja od 0,05, postoji dovoljno statističkog osnova da se odbije nulta hipoteza, što ukazuje da su raspodele verovatno različite (Virtanen et al., 2020). Takođe, pored Kolmogorov-Smirnov testa, poželjno je uraditi vizuelni prikaz originalnih i sintetičkih podataka, kao i analizirati osnovne parametre deskriptivne

statistike (srednja vrednost, medijana, modalite t, koeficijent asimetrije, koeficijent spljoštenosti i dr.) radi provere poklapanja.

2.1.2.3. Podela na trening i test set podataka

Za modele mašinskog učenja karakteristično je da se "uče" odnosno treniraju na jednom setu podataka, da bi se zatim testirali i/ili validirali na drugom setu podataka. Zbog toga je potrebno uvesti tri nova termina: trening set, test set i validacioni set podataka. Trening set podataka predstavlja skup podataka koji algoritam mašinskog učenja koristi da "nauči" povezanost između atributa i ciljne promenljive. U literaturi se često nalaze različite podele na trening i test set podataka, pri čemu se obično navodi da trening set treba da sadrži 80% celokupnog seta podataka, dok test set sadrži 20%. Kao i u prethodnom primeru koji se odnosi na količinu podataka, podela između trening i test seta podataka ne može biti striktno definisana, te istraživač može koristiti različite veličine trening i test setova, ukoliko to dozvoljava ukupna količina podataka. Takođe, treba spomenuti i validacioni set podataka, koji predstavlja treći skup podataka koji nije uključen u trening modela. Model se "uči" na trening setu i testira na test setu, dok se koristi za optimizaciju hiperparametara. Validacioni set podataka je odvojen na početku i koristi se za testiranje modela na prethodno neviđenim podacima, tek kada su hiperparametri određeni.

U praktičnim uslovima, validacioni set nije uvek korišćen, ali ako podaci dozvoljavaju takvu podelu, predstavlja dobar način za proveru modela. Važno je napomenuti da postoji potencijalna greška prilikom određivanja trening i test seta. Podaci u trening setu ne bi trebali biti isti kao podaci u test setu. Drugim rečima, model se trenira na jednom skupu podataka, ali ne bi trebao da se testira na tom istom skupu ili na podskupu trening seta[1]. Ova situacija opisuje pojavu curenja podataka (eng. *Data leakage*), što predstavlja problem i grešku prilikom modelovanja. Curenje podataka obično se ispoljava kroz veoma visoke vrednosti mera kvaliteta modela (preciznost, tačnost i dr.), čime dolazi do precenjivanja sposobnosti modela. Ukoliko dođe do curenja podataka na test setu, validacioni test treba da pokaže niže ili više vrednosti mera kvaliteta modela u odnosu na one dobijene na test setu, pod uslovom da podaci za validaciju nisu deo trening seta. Zbog toga je najbolje, kao što je prikazano na slici 1, u

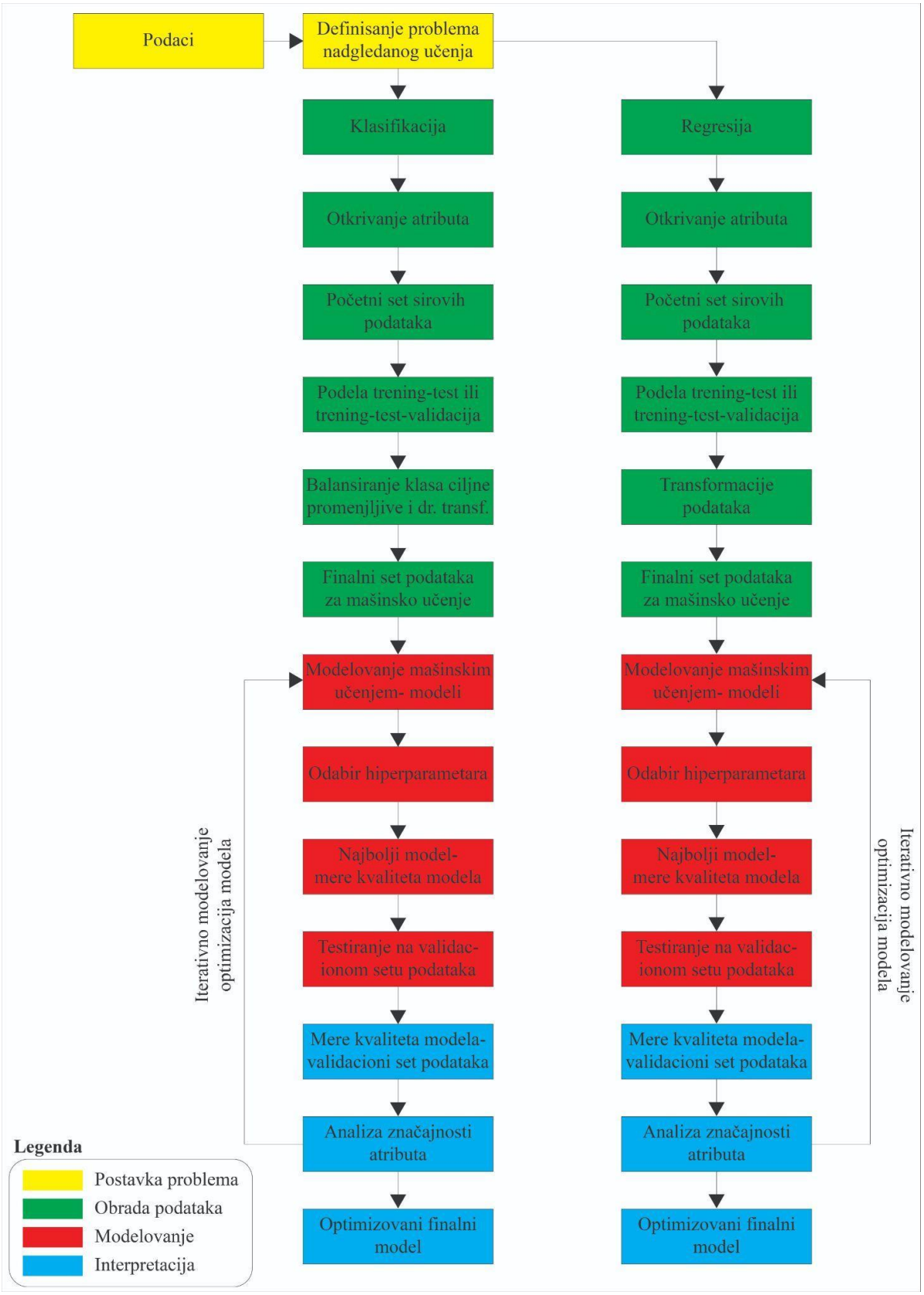
[1] U ovoj disertaciji biće prikazan primer gde se koristi trening set podataka i za testiranje modela. Uz odgovarajuće pristrastnosti i pretpostavke, ovakav pristup je uspešno primenjen.

početku odvojiti trening, test, ili trening- test- validacioni set podataka, i potvrditi da nijedan od tih setova ne sadrži podatke iz drugih setova.

2.1.2.4. Balansiranje klasa i transformacije podataka

Prva razlika prikazana na slici 1 između klasifikacije i regresije (izuzev tipa podataka ciljne promenljive) odnosi se na transformacije podataka pre modelovanja. U slučaju klasifikacije, uobičajena transformacija je balansiranje klasa ciljne promenljive u trening setu podataka. U slučaju binarne klasifikacije, odnos klasa u trening setu podataka ciljne promenljive treba da bude što bliži 50- 50%. Ukoliko je odnos klasa drugačiji, što je često slučaj, potrebno je primeniti odgovarajuće transformacije. Ako kvantitet podataka i vrsta modelovanja to dozvoljavaju, moguće je izvršiti jednostavnu transformaciju nasumičnog poduzorkovanja (eng. *Random undersampling*) na klasu koja je prezastupljena u podacima (Batista et al., 2004). Nasumično poduzorkovanje, u tom slučaju, nasumično uklanja određeni broj podataka ciljne promenljive (i atributa) dok obe klase ne postanu ravnomerno zastupljene u trening setu podataka (Hasanin & Khoshgoftaar, 2018; Saripuddin et al., 2021). Sa druge strane, ako kvantitet podataka i vrsta modelovanja ne dozvoljavaju primenu nasumičnog poduzorkovanja, mogu se koristiti druge tehnike, kao što je preuzorkovanje manjinske klase (eng. *Synthetic Minority Oversampling Technique- SMOTE*). U slučaju regresije, transformacije podataka mogu obuhvatiti logaritamsku transformaciju (logaritamska vrednost ciljne promenljive), transformaciju pomoću prirodnog korena (kvadratni, kubni ili drugi koren ciljne promenljive), diferencijaciju podataka radi dobijanja stacionarnosti ukoliko su podaci predstavljeni kao vremenska serija, itd. Detaljniji prikaz transformacija podataka u slučaju regresije biće dat u poglavlju koje se bavi modelovanjem vremenskih serija.

Nakon završetka transformacija podataka i dobijanja finalnih setova podataka koji služe kao osnova za modelovanje, naredni korak je samo modelovanje, odnosno odabir modela mašinskog učenja i njegova optimizacija kroz hiperparametre. U nastavku će biti prikazani modeli: Slučajne šume (eng. *Random Forest*), stabla odlučivanja (eng. *Decision trees*), ekstremno povećanje gradijenta (eng. *Extreme gradient boosting*), K-najbližih suseda (eng. *K-nearest neighbors*), kao i biblioteka niskog kodiranja Pajkaret (eng. *PyCaret*), koja sadrži ukupno 15 različitih modela mašinskog učenja.



Slika 1. Radni tok uopštenog modelovanja mašinskog učenja (klasifikacija i regresija)

2.1.2.5. Model stabla odlučivanja (eng. *Decision tree*)

Stabla odlučivanja (eng. *Decision trees*) predstavljaju vrlo jednostavne modele koji se koriste za klasifikaciju i regresiju. Stabla odlučivanja su preteča slučajnih šuma (eng. *Random Forest*) i kao takva, ne nalaze često primenu, jer model slučajnih šuma rešava mane koje se javljaju kod modela stabla odlučivanja. Sa druge strane, modeli stabla odlučivanja su vrlo interpretabilni i kao takvi predstavljaju odličnu polaznu tačku za prihvatanje koncepta mašinskog učenja, kao i za relativno male setove podataka gde je potrebno detaljno pratiti tok rada modela.

Stabla odlučivanja se sastoje od korena (eng. *Root node*) koji sadrži informacije o svim atributima i ciljnoj promenljivoj. Iz korena se definišu grane i čvorne tačke na osnovu podele izabranog atributa, na način da ciljna promenljiva bude što homogenija. Homogenost ciljne promenljive definiše se kao preovlađivanje jedne klase u ciljnoj promenljivoj nakon podele. Podela se vrši korišćenjem različitih indeksa, pri čemu je najčešće korišćen Đinijev indeks (eng. *Gini index*), a mogu se koristiti i entropija ili varijansa. Algoritam stabla odlučivanja vrši podele atributa do trenutka kada u listu grane (eng. *Terminal node*) preovlada što bliže jedna klasa, ili u slučaju postojanja druge klase, predikcija se vrši na osnovu glasanja, gde većina predstavlja odabranu predikciju modela. Hiperparametri stabla odlučivanja mogu uključivati maksimalnu dubinu stabla ili minimalni broj podataka u čvoru, pri čemu je faktor zaustavljanja modela definisan od strane korisnika. Drugim rečima, model stabla odlučivanja deli podatke prema atributima na način da se vrši homogenizacija ciljne promenljive, tako da u listovima stabla preovladava jedna klasa. Prilikom testiranja modela, novi atributi koje model nije video propuštaju se kroz stablo odlučivanja koje je napravljeno tokom treninga modela, što omogućava dobijanje predikcija na podacima koje model nije prethodno video.

Prednosti stabla odlučivanja ogledaju se u njegovoj jednostavnosti, interpretabilnosti i brzini dobijanja rezultata, jer model može na vrlo jednostavan način pružiti istraživaču željene predikcije na način koji je lako pratiti i objasniti. Sa druge strane, stabla odlučivanja nemaju dovoljno kapaciteta za složenije skupove podataka gde je odnos između atributa i ciljne promenljive kompleksan, sklona su pretprilagođavanju (eng. *Overfitting*), a kod velikih skupova podataka interpretabilnost može postati kompleksnija.

2.1.2.6. Model slučajnih šuma (eng. *Random Forest*)

Model slučajnih šuma (eng. *Random forest*) predstavlja jedan od najkorišćenijih modela mašinskog učenja opšte namene za klasifikaciju i regresiju. Model slučajnih šuma prvi put je prikazan od strane Brejmana (Breiman, 2001) kao ansambl stabala odlučivanja (skup više stabala odlučivanja), čime su prevaziđene mane koje imaju pojedinačna stabla odlučivanja. Od trenutka prvog prikazivanja modela 2001. godine, model slučajnih šuma našao je primenu u svim delovima nauke i privrede i trenutno je prisutan u gotovo svakom kompjuterskom paketu razvijenom za potrebe mašinskog učenja (npr. JASP, Weka itd.).

Model slučajnih šuma koristi metodu ponovnog uzorkovanja sa zamenom (eng. *Bootstrap with replacement*) na originalnom skupu podataka, uzimajući oko dve trećine podataka, koji se zatim dele na određeni broj podgrupa. Za svaku podgrupu podataka primenjuje se jedno stablo odlučivanja, pri čemu se ne koristi ceo set atributa, već samo određeni broj. Na primer, jedno stablo odlučivanja može koristiti atribute 1, 5 i 9, dok drugo stablo, stvoreno za drugi uzorak podataka, koristi atribute 2, 4 i 7. Kao i u prethodnim primerima, stablo odlučivanja koristi određeni indeks za podelu prema atributima, kako bi ciljna promenljiva nakon podele bila što homogenija. Nakon određenog broja podela, kada dalje deljenje podataka ne bi imalo smisla ili nakon dostizanja maksimalnog dozvoljenog broja podela, model se zaustavlja i novi podaci se dovode. Novi podaci su podaci iz originalnog skupa koji nisu nasumično izabrani za grupe podataka iz kojih je stablo odlučivanja stvoreno (oko jedne trećine podataka), a nazivaju se „*out-of-bag* uzorcima“. Potreba za novim podacima javlja se u svrhu računanja greške modela, poznate kao „*out-of-bag error“* (eng. *OOB error*), odnosno za internu validaciju modela. Predikcije se generišu na osnovu većeg broja prognoza modela, pri čemu se u slučaju klasifikacije vrši izbor finalne predikcije glasanjem (eng. *Voting*), dok se u slučaju regresije finalna predikcija vrši osrednjavanjem.

Prednosti modela slučajnih šuma ogledaju se u tome što, zahvaljujući korišćenju većeg broja stabala odlučivanja na različitim podskupovima podataka, gde svaki podskup koristi drugačiju kombinaciju atributa za zajedničku ciljnu promenljivu, greške su slabije korelisane za svaki podskup. Takođe, prilikom generisanja predikcija, model stvara veći broj predikcija (jednak broju podskupova podataka), koje se u slučaju klasifikacije obračunavaju kao glasanje, a u slučaju regresije kao osrednjavanje. Ovaj pristup generisanju predikcija omogućio je modelu

slučajnih šuma da postane jedan od najčešće primenjivanih modela. Sa druge strane, interpretabilnost modela slučajnih šuma, pošto predstavljaju ansambl stabala, nije moguća.

2.1.2.7. Model ekstremnog povećanja gradijenta (eng. *Extreme Gradient Boosting*)

Model ekstremnog povećanja gradijenta (eng. *Extreme Gradient Boosting- XGB*), kao i model slučajnih šuma, predstavlja ansambl stabala odlučivanja. Model ekstremnog povećanja gradijenta prikazan je prvi put 2016. godine (Chen & Guestrin, 2016) i postao je jedan od najčešće primenjivanih modela mašinskog učenja u različitim oblastima.

Razlika između ekstremnog povećanja gradijenta i slučajnih šuma ogleda se u radnom toku modela. Model slučajnih šuma paralelno gradi stabla odlučivanja na podskupovima originalnog skupa podataka, dok model ekstremnog povećanja gradijenta vrši predikcije tako što u inicijalnom koraku prognozira sve vrednosti da budu ista klasa, tj. jedna klasa ciljne promenljive. Nakon toga, sračunava se funkcija gubitka (eng. *Loss function*) i izračunavaju prvi i drugi izvodi za svaki podatak. Ovi izvodi se koriste prilikom izgradnje prvog stabla odlučivanja, kako bi se odredilo prema kom atributu bi se vršila podela na način da se funkcija gubitka smanji. Nakon izgradnje prvog stabla odlučivanja, predikcije se resetuju prema predikcijama tog stabla. Prethodno opisani koraci se iterativno ponavljaju, tako da svako naredno stablo odlučivanja koriguje greške prethodnog stabla, tj. greške koje prethodno stablo nije ispravilo u odnosu na inicijalne prognoze. Iterativno ponavljanje se vrši dok smanjenje funkcije gubitka ne postane zanemarljivo. Nakon što su sva stabla odlučivanja određena, model koristi ponderisanu sumu za finalnu predikciju, za razliku od modela slučajnih šuma koji koristi glasanje (klasifikacija) ili osrednjavanje (regresija).

2.1.2.8. Model K- najbližih suseda (eng. *K- Nearest Neighbors*)

Model K-najbližih suseda (eng. *K- nearest neighbors*) je prvi model koji je prikazan u ovom sklopu, a koji nije zasnovan na stablima odlučivanja. Model K- najbližih suseda jedan je od najjednostavnijih modela mašinskog učenja, koji se češće koristi za klasifikaciju nego za regresiju. Prognoza sa metodom K- najbližih suseda vrši se tako što algoritam za podatak iz test seta podataka pronalazi najsličnije podatke prema atributima iz trening seta i pridružuje mu klasu koja se najčešće javlja među tim podacima.

U modelu K- najbližih suseda ne postoji tradicionalno treniranje modela, kao što je to slučaj sa modelima slučajnih šuma ili ekstremnog povećanja gradijenta. Umesto toga, vrši se poređenje novog podatka sa sličnim podacima u trening setu. Takođe, pošto ne postoji trening modela, ne vrši se ni optimizacija modela kao kod modela ekstremnog povećanja gradijenta. Interpretabilnost modela K- najbližih suseda nije moguća, baš kao ni kod modela slučajnih šuma. Pored toga, model K- najbližih suseda pokazuje generalno lošije rezultate u prognoziranju u poređenju sa drugim modelima, ali može dati zadovoljavajuće rezultate u nekim slučajevima, iako je jednostavan. Takođe, model K- najbližih suseda je osetljiv na vrednosti podataka, pa je preporučljivo skalirati podatke (normalizacija ili standardizacija) pre njegove primene. Potrebno je naglasiti da model K- najbližih suseda ima lošije performanse u slučaju velikog broja atributa (eng. *The curse of dimensionality*), te je potrebno ili izdvojiti relevantne atribute ili primeniti redukciju dimenzionalnosti atributa (eng. *Dimensionality reduction*).

2.1.2.9. Biblioteka PyCaret

Prethodno prikazani modeli mašinskog učenja predstavljaju modele koji su vrlo često u upotrebi i dobro poznati u literaturi. Sa druge strane, postoje biblioteke koje sadrže skupove modela, a koje su napravljene za brzo treniranje i laku implementaciju modela u nastavku procesa. PyCaret biblioteka za programski jezik Pajton (eng. *Python*) predstavlja biblioteku niskog koda (eng. *low- code library*) koja je razvijena za potrebe mašinskog učenja. Jednostavna implementacija većeg broja modela predstavlja značajnu prednost ove biblioteke, zbog čega je izabrana za korišćenje. Izlaz nakon treniranja modela prikazuje mere kvaliteta modela, koje se dobijaju nakon kros-validacije za svaki model. Pored modela slučajnih šuma, ekstremnog povećanja gradijenta, stabala odlučivanja i K-najbližih suseda, u sklopu ove biblioteke nalazi se još 12 modela, među kojima su: *Logistička regresija, Ridge klasifikator, Linearno diskriminantna analiza, Naivni Bajes, CatBoost klasifikator, Klasifikator gradijentnog pojačanja, Ada Boost klasifikator, Extra Trees klasifikator, Kvadratna diskriminantna analiza, Light Gradient Boosting Machine, Dummy klasifikator, SVM-Linearna jezgra.*

Na kraju, treba napomenuti da, pored implementacije individualnih modela ili grupa modela, kao što je slučaj sa PyCaret bibliotekom u izabranom programskom jeziku, moguće je koristiti

i gotova softverska rešenja poput JASP-a, Weka i drugih, koji takođe nude veliki broj modela korisnicima na vrlo intuitivan način, opremljen sa korisničkim interfejsom.

2.1.2.10. Optimizacija hiperparametara

Optimizacija hiperparametara (eng. *Hyperparameter optimization* ili *Hyperparameter tuning*) predstavlja jedan od ključnih koraka prilikom treniranja modela. U zavisnosti od implementiranog algoritma, različiti hiperparametri mogu biti optimizovani. U praksi postoje dva načina koji se mogu koristiti za pronalaženje najoptimalnijih hiperparametara: potraživanje hiperparametara prema gridu (eng. *Grid Search*) i nasumično (eng. *Random Search*).

Potraživanje hiperparametara prema gridu predstavlja najtemeljniji način potraživanja hiperparametara datog modela. Praktično, ovaj metod zahteva definisanje početnih parametara, krajnjih parametara i intervala po kojem se parametri pretražuju. Na primer, za algoritam slučajne šume, potraživanje prema gridu bilo bi definisano kao početni broj stabala X, krajnji broj stabala Y i interval Z. Parametre X, Y i Z moguće je definisati prema potrebama istraživanja. Na primer, može biti potrebno pretražiti prostor hiperparametara koji počinje sa 50 stabala, završava sa 500 stabala, a interval je 25 stabala. U tom slučaju, model će pretražiti ukupno 19 modela. Kada se koristi potraživanje hiperparametara prema gridu, ukupan broj modela može se dobiti pomoću sledeće jednačine:

$$ukupan\ broj\ modela = \left(\frac{krajnji\ hiperparametar - početni\ hiperparametar)}{interval}\right) + 1. \qquad (1)$$

U slučaju nasumičnog potraživanja hiperparametara potraživanje u definisanom prostoru nasumično odabira vrednosti datog hiperparametra za testiranje. Ukupan broj modela za testiranje zavisi od ukupnog broja modela koji istraživač definiše. Drugim rečima, definišu se početni i krajnji interval, kao i ukupan broj modela. Potraživanje se vrši nasumičnim izvlačenjem vrednosti u datom regionu, gde je ukupan broj modela definisan od strane istraživača.

U praktičnim uslovima nije moguće dati preciznu definiciju koja metoda je bolja; sve zavisi od seta podataka, cilja istraživanja i računarskih resursa. Sa druge strane, potraživanje prema gridu, ukoliko je prostor hiperparametara ili set podataka veliki, ili interval mali, može

zahtevati veliko računarsko vreme i resurse. U tom slučaju, nasumično potraživanje je bolja opcija.

Prethodno su prikazani različiti modeli mašinskog učenja, ali nije bilo reči o optimizaciji hiperparametara, niti o hiperparametrima koje različiti modeli pružaju za optimizaciju. U slučaju stabla odlučivanja, kao hiperparametar za optimizaciju može se koristiti kriterijum zaustavljanja, dok je kod modela slučajnih šuma broj stabala jedan od glavnih hiperparametara. Broj stabala kod slučajnih šuma je interesantan hiperparametar jer veće vrednosti tog parametra obično poboljšavaju sposobnost modela u zadatom zadatku, ali na uštrb računarskog vremena (Nikolić & Zečević, 2019). Kod modela ekstremnog povećanja gradijenta, broj stabala je takođe jedan od glavnih hiperparametara, kao i stopa učenja (eng. *Learning rate*). Vrednosti stope učenja obično se kreću od 0 do 1, pri čemu se najčešće koriste male vrednosti, npr. od 0,01 do 0,3. U modelu K- najbližih suseda ključni hiperparametar predstavljen je brojem najbližih suseda ili K- vrednošću. U biblioteci *PyCaret*, optimizacija hiperparametara je već definisana u pozadini algoritma, te se prilikom rangiranja modela prikazuju optimizovane vrednosti tih modela. Takođe, prilikom testiranja datog modela moguće je koristiti njegovu najoptimizovaniju verziju, čime se olakšava proces modelovanja.

Na kraju, treba naglasiti da odabir odgovarajućeg hiperparametra u velikoj meri utiče na tačnost modela, te je od velikog značaja posvetiti posebnu pažnju ovom koraku prilikom modelovanja. Kod kompleksnijih modela, optimizacija hiperparametara je složenija. Primer za to može biti grupa modela kao što su stabla odlučivanja, slučajne šume i ekstremno povećanje gradijenta. Optimizacija hiperparametara kod stabla odlučivanja je najjednostavnija, dok slučajne šume imaju više hiperparametara, od kojih je glavni broj stabala. Kod ekstremnog povećanja gradijenta, dva glavna parametra su broj stabala i stopa učenja. Ovaj primer pokazuje da, kako se povećava kompleksnost modela, tako raste i kompleksnost odabira odgovarajućih hiperparametara. Na krajnjem delu spektra nalaze se metode veštačke inteligencije, kao što su neuronske mreže, koje imaju veći broj hiperparametara za optimizaciju, zbog čega je kod njih posebno važno pažljivo odabrati sve hiperparametre. Optimizacija hiperparametara je ista u slučaju klasifikacije i regresije.

2.1.2.11. Mere kvaliteta modela klasifikacije

Prilikom testiranja i validacije modela, atributi u test i/ili validacionom setu podataka koriste se za dobijanje vrednosti klasa na osnovu prethodno stečene spoznaje iz trening seta podataka. Test set podataka (kao i validacioni set podataka) sadrže vrednosti ciljne promenljive, ali te vrednosti se ne koriste za obučavanje modela, već za kvantifikaciju mere kvaliteta modela. Mere kvaliteta modela u slučaju klasifikacije mogu biti: matrica konfuzije (eng. *Confusion matrix*), tačnost (eng. *Accuracy*), preciznost (eng. *Precision*), osetljivost ili stopa tačno pozitivnih instanci (eng. *Sensitivity* ili *true positive rate*), stopa lažno pozitivnih instanci (eng. *False positive rate*), stopa lažnih otkrića (eng. *False discovery rate*), Metjuzov koeficijent korelacije (eng. *Matthews correlation coefficient*), površina ispod AUC krive (eng. *Area under the Reciever Operating Characteristic (ROC) curve*), F1-mera (eng. *F1-Score*), negativna prediktivna vrednost (eng. *Negative predictive value*), stopa tačno negativnih instanci (eng. *Specificity* ili *True negative rate*), stopa lažno negativnih instanci (eng. *False negative rate*) i statistička jednakost (eng. *Statistical parity*).

Matrica konfuzije predstavlja prvi vid evaluacije modela nakon dobijanja rezultata sa test ili validacionog seta podataka i takođe je osnova za računanje drugih navedenih parametara kvaliteta evaluacije modela. Matrica konfuzije se dobija poređenjem stvarnih vrednosti ciljne promenljive sa izlazom ciljne promenljive iz modela. U slučaju binarne klasifikacije, matricu konfuzije moguće je prikazati primerom iz tabele 1.

Tabela 1. Matrica konfuzije u slučaju binarne klasifikacije

		Prognizirane vrednosti	
		Pozitivno	**Negativno**
Istinite vrednosti	**Pozitivno**	TP	FN
	Negativno	FP	TN

Matrica konfuzije u slučaju binarne klasifikacije data je u tabeli 1, prikazuje četiri mogućnosti za klasifikaciju izlaza modela ciljne promenljive. Tačno pozitivna (TP) vrednost predstavlja instancu koja je okarakterisana kao pozitivna u test setu podataka i tačno je klasifikovana od strane modela. Tačno negativna (TN) vrednost predstavlja instancu koja je okarakterisana kao negativna u test setu podataka i takođe je tačno klasifikovana od strane modela. Sa druge strane, lažno negativna (FN) predstavlja instancu koja je okarakterisana kao pozitivna u test setu

podataka, ali je prognozirana kao negativna. Lažno pozitivna (FP) predstavlja instancu koja je okarakterisana kao negativna u test setu podataka, a klasifikovana je od strane modela kao pozitivna vrednost. U ovom primeru, moguće je zameniti pozitivne i negativne instance sa vrednostima 1 i 0 (binarne vrednosti), odgovorima "Da" ili "Ne" itd. U slučaju idealnog klasifikatora, koji u praksi ne bi trebalo da postoji, vrednosti TP i TN bile bi jednake pozitivnim i negativnim vrednostima u test setu podataka, dok bi FP i FN vrednosti bile jednake 0. U praktičnim uslovima, takav ishod je vrlo retko slučaj i obično predstavlja indikaciju da je došlo do curenja podataka, ili da u atributima postoji vrednost koja predstavlja ciljnu promenljivu (takođe vid curenja podataka).

Parametar tačnosti se računa iz matrice konfuzije na sledeći način:

$$tačnost = \frac{(TP+TN)}{(TP+TN+FP+FN)}. \quad (2)$$

Moguće je dati i formalnu definiciju parametra tačnosti, koja predstavlja odnos tačno klasifikovanih instanci modela naspram svih klasifikovanih instanci. Vrednosti tačnosti se kreću u rasponu od 0 do 1, gde vrednost 1 predstavlja idealan klasifikator.

Parametar preciznosti predstavlja odnos TP i FP vrednosti, odnosno:

$$preciznost = \frac{TP}{TP+FP}. \quad (3)$$

Stopa tačno pozitivnih instanci (osetljivost) i stopa lažno pozitivnih instanci se definišu kao:

$$osetljivost = \frac{TP}{TP+FN}, \quad (4)$$

$$stopa\ lažno\ pozitivnih\ instanci = \frac{FP}{FP+TN}. \quad (5)$$

Parametar osetljivosti prikazuje odnos tačno klasifikovanih pozitivnih instanci u odnosu na ukupni broj pozitivnih instanci, dok parametar stope lažno pozitivnih instanci prikazuje odnos broja instanci koje su netačno klasifikovane kao pozitivne (lažno pozitivne) u odnosu na ukupni broj instanci koje su negativne.

Stopa lažnih otkrića data je kao:

$$stopa\ lažnih\ otkrića = \frac{FP}{TP+FP}\,. \tag{6}$$

Stopa lažnih otkrića prikazuje broj lažno pozitivnih instanci prema ukupnom broju pozitivnih instanci.

Metjuzov koeficijent korelacije (MCC) prikazan je kao:

$$MCC = \frac{(TP\times TN - FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\,. \tag{7}$$

Metjuzov koeficijent korelacije smatra se vrlo korisnom merom ocene kvaliteta modela, pošto koristi sve vrednosti matrice konfuzije i zavisi od toga kako model prikazuje uspešne klasifikacije u sve četiri kategorije. U slučaju vrlo nebalansiranih setova podataka, MCC se smatra boljom merom ocene kvaliteta modela u odnosu na preciznost, jer preciznost može prikazati prividno povećane vrednosti u tom slučaju (Chicco & Jurman, 2020).

Površina ispod ROC (u nastavku AUC) krive predstavlja jedan od najčešće korišćenih individualnih parametara ocene kvaliteta modela. Primena površine ispod AUC krive koristi se kako bi se ocenila sposobnost modela da pravi razliku između dve klase (u slučaju binarne ciljne promenljive). Vrednosti AUC se nalaze između 0 i 1, dok se najčešće sreću vrednosti od 0,5 do 1, gde vrednost 0,5 predstavlja potpunu nesposobnost modela da razlikuje između dve klase (nasumično pogađanje klasa), dok vrednost 1 predstavlja idealan klasifikator koji u potpunosti pravi razliku između klasa. U praktičnim uslovima, vrednosti veće od 0,8 se smatraju zadovoljavajućim (Çorbacıoğlu & Aksel, 2023).

F1-mera je jedan od najkorišćenijih parametara ocene kvaliteta modela, pored površine ispod AUC krive, i definisana je kao:

$$F1 = 2\times\frac{(preciznost\times osetljivost)}{(preciznost+osetljivost)}\,. \tag{8}$$

F1-mera predstavlja harmonijsku srednju vrednost između preciznosti i osetljivosti i, kao takva, često se koristi u slučajevima velike razlike između odnosa klasa. Takođe, predstavlja bolju meru ocene kvaliteta modela u odnosu na preciznost (Joshi, 2002; Hossin & Sulaimani, 2015), a u daljem radu će se najviše oslanjati na F1- meru.

Negativna prediktivna vrednost računa se prema:

$$NPV = \frac{TN}{(TN+FN)}. \tag{9}$$

Negativna prediktivna vrednost predstavlja odnos tačno klasifikovanih negativnih instanca u odnosu na celokupni broj negativnih instanca u setu podataka.

Stopa lažno negativnih ili specifičnost računa se kao:

$$specifičnost = \frac{TN}{TN+FP}. \tag{10}$$

Parametar stope lažno negativnih vrednosti ili specifičnost predstavlja sličan parametar osetljivosti, ali za drugu (negativnu) klasu. Specifičnost izražava broj tačno negativnih instanci u odnosu na sve klasifikovane negativne instance.

Stopa lažno negativnih vrednosti je definisana kao:

$$stopa\ lažno\ negativnih\ vrednosti = \frac{FN}{FN+TP}. \tag{11}$$

Stopa lažno negativnih vrednosti predstavlja odnos pozitivnih instanci koje su klasifikovane kao negativne u odnosu na sve pozitivne instance.

Statistička jednakost predstavlja jednu od najjednostavnijih mera ocene kvaliteta modela, jer prikazuje odnos između klasa u test setu podataka dobijenih od izlaza modela. Takva vrednost treba da bude približno ista kao i stvarni odnos klasa u test setu podataka u idealnom slučaju.

Na kraju, potrebno je napomenuti da je moguće sve prethodno prikazane mere kvaliteta modela sračunati za obe klase (u slučaju binarne ciljne promenljive) ili za više klasa u slučaju ciljne

promenljive sa više od dve klase, čime se dobija detaljniji prikaz ocene kvaliteta modela. U ovoj disertaciji, za obe klase (najčešće) korišćen je parametar F1- mere.

2.1.2.12. Mere kvaliteta modela regresije

Za razliku od prethodno prikazanih mera kvaliteta modela klasifikacije, koje prikazuju razliku između oznaka klasifikovane i stvarne ciljne promenljive, mere kvaliteta modela regresije prikazuju razliku između prognozirane i stvarne brojčane vrednosti. U mere kvaliteta modela regresije ubrajaju se: apsolutna greška (eng. *Absolute error*), osrednjena apsolutna greška (eng. *Mean absolute error*), relativna greška (eng. *Absolute percentage error*), osrednjena relativna greška (eng. *Mean absolute percentage error*), koren srednje kvadratnog odstupanja (eng. *Root mean square error*) i koeficijent determinacije (eng. *Coefficient of determination*).

Pored prethodno prikazanih mera kvaliteta modela regresije, postoje i druge izvedene mere kvaliteta modela regresije. Međutim, treba napomenuti da, u većini slučajeva, mere poput osrednjene apsolutne greške i osrednjene relativne greške predstavljaju dovoljno dobre, lako izvedene i interpretabilne mere kvaliteta modela regresije. Moguće je prikazati i druge statistike vezane za odstupanja modela regresije, kao što su minimalna i maksimalna apsolutna i relativna greška, ukoliko je potrebno imati informaciju o ekstremnim odstupanjima modela regresije.

Apsolutna greška (AE) data je izrazom:

$$AE = \left| y_{prognozirano} - y_{istinito} \right| . \tag{12}$$

Osrednjena apsolutna greška (MAE) može se prikazati jednačinom:

$$MAE = \frac{\sum_{i=1}^{n} \left| y_{prognozirano} - y_{istinito} \right|}{n}, \tag{13}$$

gde n predstavlja broj podataka ili prognozirajući horizont.

Relativna greška (APE) i osrednjena relativna greška (MAPE) se mogu prikazati izrazima:

$$APE = \frac{\left| y_{prognozirano} - y_{istinito} \right|}{y_{istinito}} \quad \text{i} \tag{14}$$

$$MAPE = 100 \times \frac{1}{n}\sum_{i=1}^{n} \frac{|y_{prognozirano} - y_{istinito}|}{y_{istinito}}. \quad (15)$$

Koren srednje kvadratnog odstupanja (RMSE) predstavlja često korišćen parametar u geofizici za kvantifikaciju odstupanja modela prilikom inverzije. Takođe, koren srednje kvadratnog odstupanja ima svoju primenu i prilikom kvantifikacije odstupanja modela regresije, i to je izraženo sledećim izrazom:

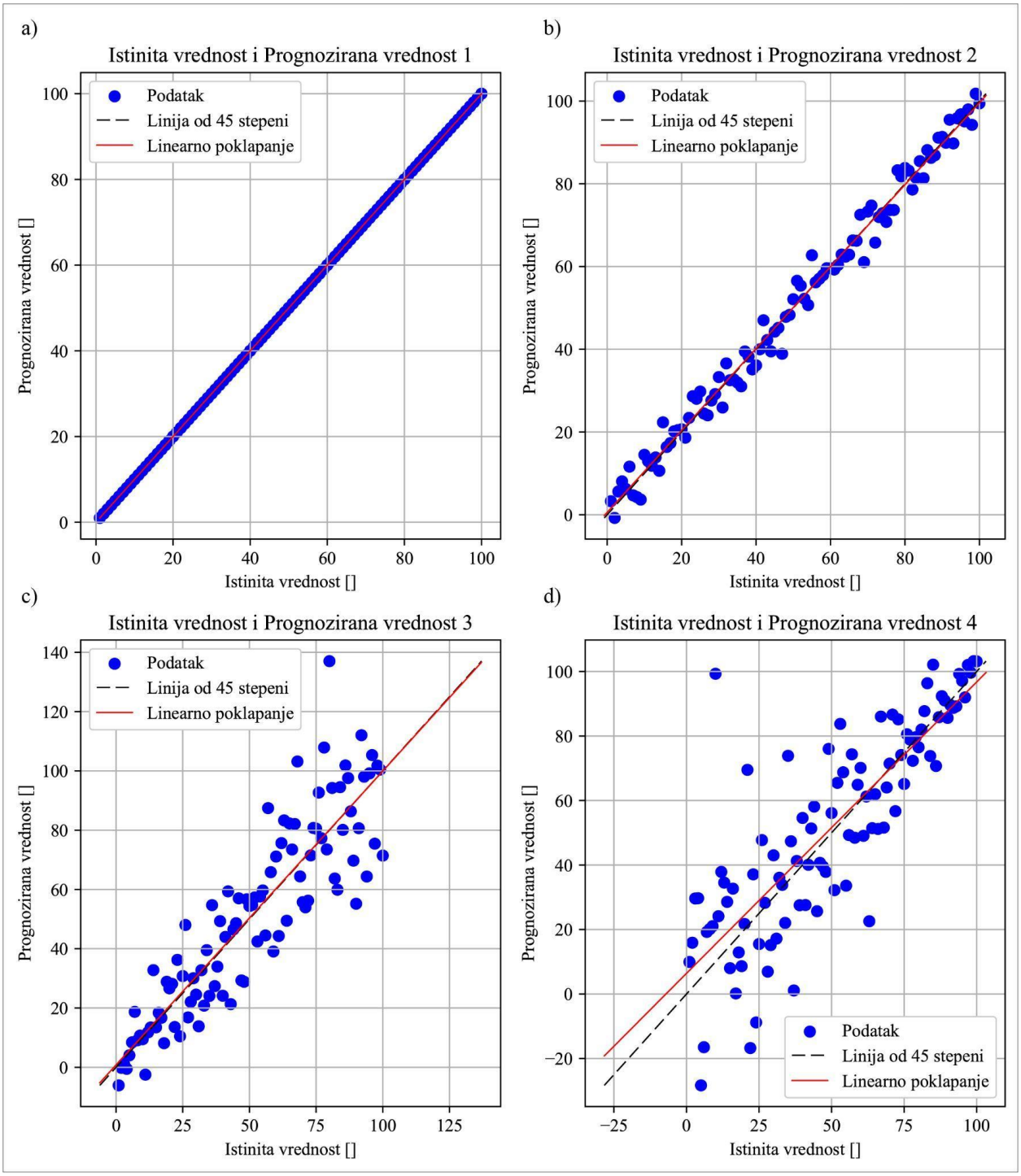
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{istinito} - y_{prognozirano})^2}{n}}. \quad (16)$$

Takođe je moguće dobiti i srednje kvadratno odstupanje (eng. *Mean square error- MSE*) zanemarivanjem kvadratnog korena u izrazu (16) što predstavlja još jednu primenjenu meru ocene kvaliteta modela regresije.

Pored prethodno prikazanih brojčanih vrednosti koje kvantifikuju tačnost modela regresije, postoje i vizuelne metode prikazivanja odstupanja modela. Jedna vizuelna metoda predstavlja prikazivanje na grafikonu prognoziranih i stvarnih vrednosti. Slika 2 ilustruje slučaj vizuelne interpretacije podataka dobijenih iz modela i stvarnih podataka[2]. Slika 2a prikazuje teorijski slučaj, gde postoji potpuno poklapanje stvarnih i prognoziranih vrednosti. U ovom slučaju, vrednosti koeficijenta determinacije predstavljene su idealnom vrednošću od 1, dok su vrednosti osrednjene apsolutne i relativne greške nepostojeće[3]. Isto tako, kriva od 45 stepeni i linearno poklapanje podataka nalaze se jedna na drugoj, što se prikazuje kod idealnih modela.

Realniji slučaj je prikazan na slici 2b, gde je vrednost koeficijenta determinacije i dalje visoka, sa vrednošću od 0,98, dok su vrednosti osrednjene apsolutne greške 2,53, a osrednjene relativne greške oko 13%. Maksimalne apsolutne i relativne greške za primer sa Slike 2b prikazane su vrednostima od 8,1 i 227%. Slučajevi gde model pokazuje određenu vrstu pristrasnosti prema nižim (Slika 2c) ili višim (Slika 2d) vrednostima se ogleda u umanjenju greške za jedan deo prognoza, a njenom uvećanju za drugi deo prognoza.

[2] Za sliku 2, u svrhe ilustracije, su korišćeni sintetički podaci.
[3] Ovakav slučaj se uglavnom ne sreće prilikom modelovanja vođenim podacima, a ukoliko se sretne može da ukazuje na curenje podataka.



Slika 2. Grafički prikaz istinitih i prognoziranih vrednosti za (a) Idealan regresioni model; (b) Realan regresioni model; (c) Model koji prikazuje pristrasnost prema nižim vrednostima i (d) Model koji prikazuje pristrastnost prema višim vrednostima

Slučaj sa Slike 2c i Slike 2d je lakše prikazati, interpretirati i uočiti pristrasnost modela prilikom grafičkog prikaza reziduala. Reziduali predstavljaju količinu odstupanja prognozirane

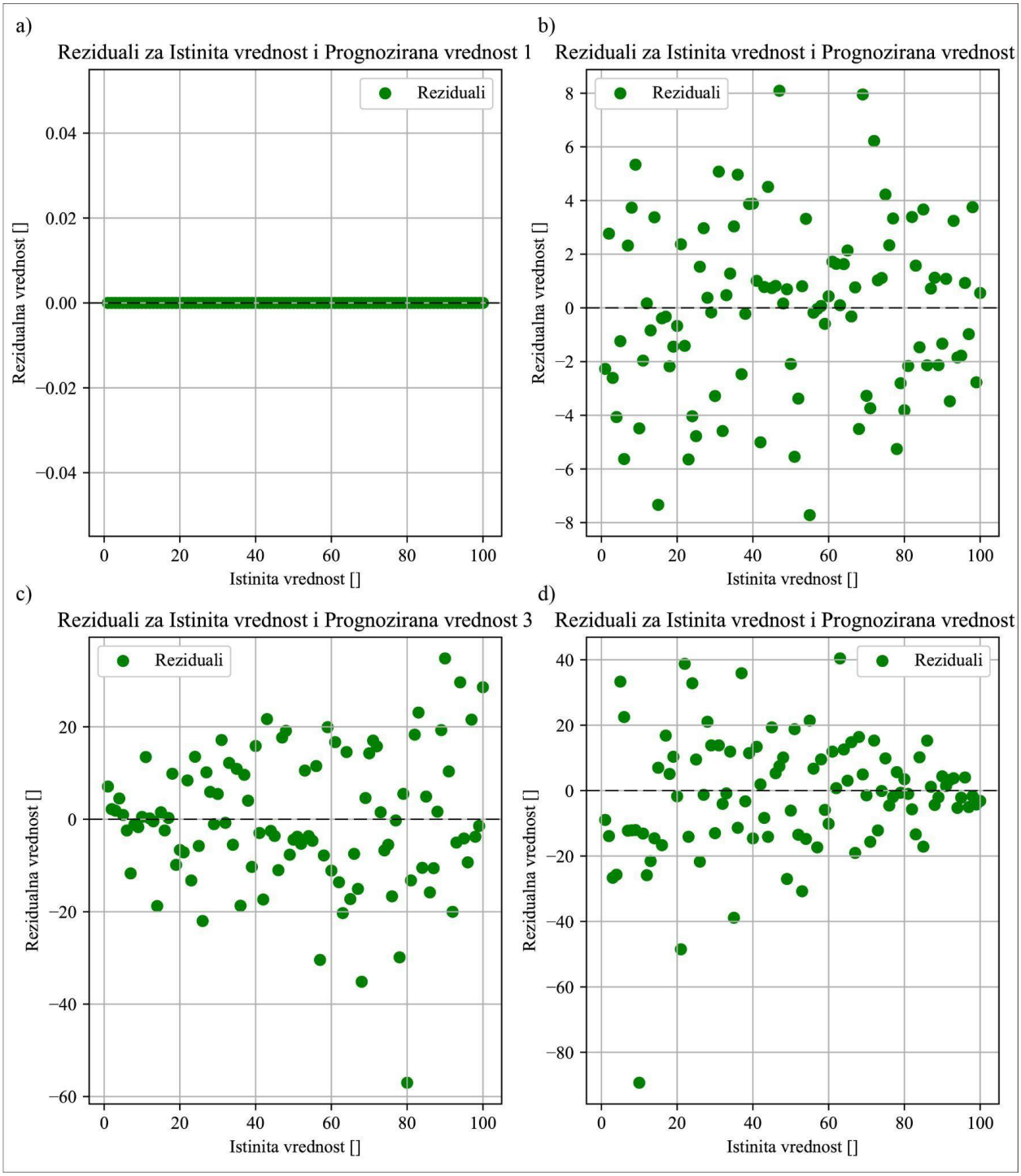
vrednosti od stvarne vrednosti. Kod idealnog regresionog modela (Slika 3a[4]), reziduali su svi jednaki nuli. U realnim slučajevima, očekuje se da reziduali budu nasumično raspoređeni na grafiku i da ne pokazuju nikakvu vrstu šablona prema kojem se ponašaju.

Slika 3c i slika 3d prikazuju pristrasnost modela, gde se na slici 3c može uočiti povećanje vrednosti reziduala sa povećanjem stvarne vrednosti, dok se na slici 3d vidi obrnuta situacija. Oba slučaja pokazuju da model ima poteškoća u prognoziranju određenih vrednosti, odnosno da u jednom slučaju prikazuje prognoze sa zadovoljavajućim odstupanjem koje se povećava sa rastom stvarnih vrednosti, ili obrnuto. Svojstvo koje opisuje prethodno prikazane reziduale naziva se homoskedastičnost u slučaju nasumičnih reziduala ili heteroskedastičnost u slučaju reziduala koji prikazuju određeni šablon (povećanje ili smanjenje reziduala sa povećanjem ili smanjenjem stvarne vrednosti). Potrebno je napomenuti da za model koji se smatra završnim, heteroskedastičnost treba korigovati ukoliko je prisutna.

Iako su apsolutna i relativna greška, kao i njihove osrednjene i druge izvedene statističke vrednosti lako razumljive i interpretabilne, postoji određeno ograničenje prilikom njihove implementacije. Prethodni slučaj sa slike 2b prikazao je maksimalnu vrednost apsolutne greške od 8,1, dok je maksimalna vrednost relativne greške 227%. Vrednosti maksimalnih apsolutnih i relativnih grešaka ne odgovaraju istom podatku. Naime, za apsolutno odstupanje od 8,1, koje je ujedno i najveće apsolutno odstupanje, odgovara relativno odstupanje od 17% (istinita vrednost 47, dok je prognozirana 38,9). Sa druge strane, maksimalnom relativnom odstupanju od 227% odgovara apsolutno odstupanje od 2,27 (istinita vrednost 1, prognozirana 3,27).

Ovaj primer ilustruje da su apsolutna i relativna greška u funkciji reda veličine stvarne i prognozirane vrednosti. Prilikom prognoziranja relativno malih vrednosti, procentualna greška biće velika za malo apsolutno odstupanje i obrnuto. U takvim slučajevima, može biti korisno potražiti druge mere ocene kvaliteta modela koje će bolje odgovarati datom zadatku.

[4] Za sliku 3, u svrhe ilustracije, su korišćeni sintetički podaci.



Slika 3. Grafički prikaz reziduala i istinitih vrednosti za (a) Idealan regresioni model; (b) Realan regresioni model; (c) Model koji prikazuje pristrasnost prema nižim vrednostima i (d) Model koji prikazuje pristrastnost prema višim vrednostima

2.1.2.13. Analiza značajnosti atributa

Modeli kao što su slučajne šume, stabla odlučivanja, ekstremno povećanje gradijenta i drugi, prilikom završetka testiranja modela, generišu podatke o značajnosti atributa. Analiza

značajnosti atributa predstavlja gradaciju atributa korišćenih za modelovanje prema određenim parametrima koji brojčano izražavaju značajnost tih atributa. Drugim rečima, model prikazuje koji su najinformativniji atributi iz skupa svih atributa bili za modelovanje ciljne promenljive. Lako se može uvideti zašto je ovo bitan korak prilikom modelovanja, jer se najpre dobijaju informacije o uzročno- posledičnoj vezi, tj. korelaciji atributa i ciljne promenljive, a sa druge strane dobijaju se informacije o mogućoj uštedi računarskog vremena prilikom modelovanja.

Analiza značajnosti atributa se različito izražava u različitim softverskim paketima. Na primer, u softverskom paketu JASP, analiza značajnosti atributa izražena je sa osrednjenim smanjenjem tačnosti čvora (eng. *Mean decrease in node accuracy*) i ukupnim povećanjem homogenosti čvora (eng. *Total increase in node purity*), ili drugačije nazvanim Đinijev indeks. Osrednjeno smanjenje tačnosti čvora predstavlja parametar koji izražava koliko bi se smanjenje čvora u modelu ostvarilo ukoliko bi se dati atribut uklonio, dok sa druge strane, ukupno povećanje homogenosti čvora prikazuje koliko dati atribut doprinosi homogenosti datog čvora. Prilikom interpretacije oba prethodna parametra, veće vrednosti ukazuju na veću značajnost, odnosno informativnost datog atributa. Biblioteke koje su dostupne u Pajtonu uglavnom koriste Đinijev indeks ili entropiju prilikom rangiranja atributa, pri čemu nema razlike u interpretaciji značajnosti atributa, bez obzira na korišćenu metodu (veće vrednosti ukazuju na veću informativnost datog atributa).

Analiza značajnosti atributa, pored prikaza koji su atributi najinformativniji za prognoziranje ili klasifikaciju ciljne promenljive, takođe pruža potencijalnu osnovu za unapređenje modela kroz uštedu na računarskom vremenu time što se uklone oni atributi koji su se pokazali da su od manje značajnosti u sprovedenoj analizi nakon upoređivanja, tzv. neinformativni atributi.

2.1.2.14. Interpretacija modela i iterativno modelovanje

Prilikom dobijanja prvog finalnog modela, moguće je izvršiti interpretaciju celokupnog procesa modelovanja i napraviti dalje korake ka optimizaciji datog modela. Prvi finalni model se ne bi trebaolo smatrati krajnjim, jer proces modelovanja mašinskim učenjem predstavlja iterativni proces optimizacije modela, pronalaženja dodatnih atributa koji mogu doprineti modelu, kao i vršenja drugih optimizacija, poput promene metode traženja optimalnih hiperparametara, promene prostora pretraživanja hiperparametara, transformacije atributa i drugih do postizanja željene tačnosti.

Poželjno je prilikom dobijanja finalnog seta trening i test podataka izvršiti analizu faktora inflacije varijanse (eng. *Variance Inflation Factor*), koja prikazuje stepen multikolinearnosti (sličnosti) između atributa. Ovaj korak je bitan prilikom modelovanja mašinskim učenjem jer multikolinearnost između atributa predstavlja otežavajuću okolnost za model, jer iste informacije mogu biti prisutne u dva atributa, a to povećava kompleksnost i računarsko vreme. Primer za multikolinearnost između atributa najbolje se može ogledati u meteorološkim parametrima temperature i prividne temperature. Oba parametra iskazuju određeni vid temperature, ali nisu potpuno isti. Faktor inflacije varijanse u tom slučaju biće vrlo veliki za date atribute, pa je poželjno ukloniti jedan od njih. Takođe, moguće je primeniti Pirsonov (eng. *Pearson*) ili Spirmanov (eng. *Spearman*) koeficijent korelacije ukoliko su podaci podobni za tu vrstu analize, radi dobijanja korelacije između datih parametara. Pored toga, moguće je spojiti više atributa u jedan primenom analize glavnih komponenata (eng. *Principal Component Analysis*) i time dobiti jedan informativan atribut.

Potrebno je napomenuti otežavajuću okolnost nadgledanog mašinskog učenja koja se odnosi na obeležavanje skupa podataka (eng. *Data labeling*) koji se koristi za „učenje mašine“. Naime, nadgledano mašinsko učenje očekuje da za svaku vrednost atributa postoji vrednost ciljne promenljive. U određenim situacijama, priprema podataka za mašinsko učenje iziskuje od istraživača pripremu obeleženih uzoraka za model, odnosno određivanje ciljne promenljive. U nekim slučajevima, moguće je relativno jednostavno dobiti vrednosti ciljne promenljive, dok u drugim to iziskuje veliku količinu vremena istraživača ili istraživačke grupe. U daljem tekstu biće prikazan primer koji upravo ukazuje na taj nedostatak, ali će biti prikazan i način „olakšavanja“ datih nedostataka kod nadgledanog mašinskog učenja.

Kao što je prethodno spomenuto, polje mašinskog učenja, nauke o podacima, primenjene statistike i modela vođenih podacima je vrlo široko, te je predstavljanje svih metoda, tehnika, transformacija i drugih u jednom tekstu vrlo zahtevan zadatak. Obično, primena različitih tehnika obrade podataka zavisi od specifičnog slučaja, a izrada jednog sveobuhvatnog radnog toka sa prikazom svih mogućih metoda je gotovo nemoguća. Prethodno prikazani tekst imao je za cilj da prikaže uopšteni radni tok mašinskog učenja i osnovne koncepte klasifikacije, regresije i interpretacije modela. U nastavku teksta, specifičnosti vezane za svaki individualni slučaj biće prikazane u kontekstu datog slučaja.

2.2. Metode prognoziranja vremenskih serija

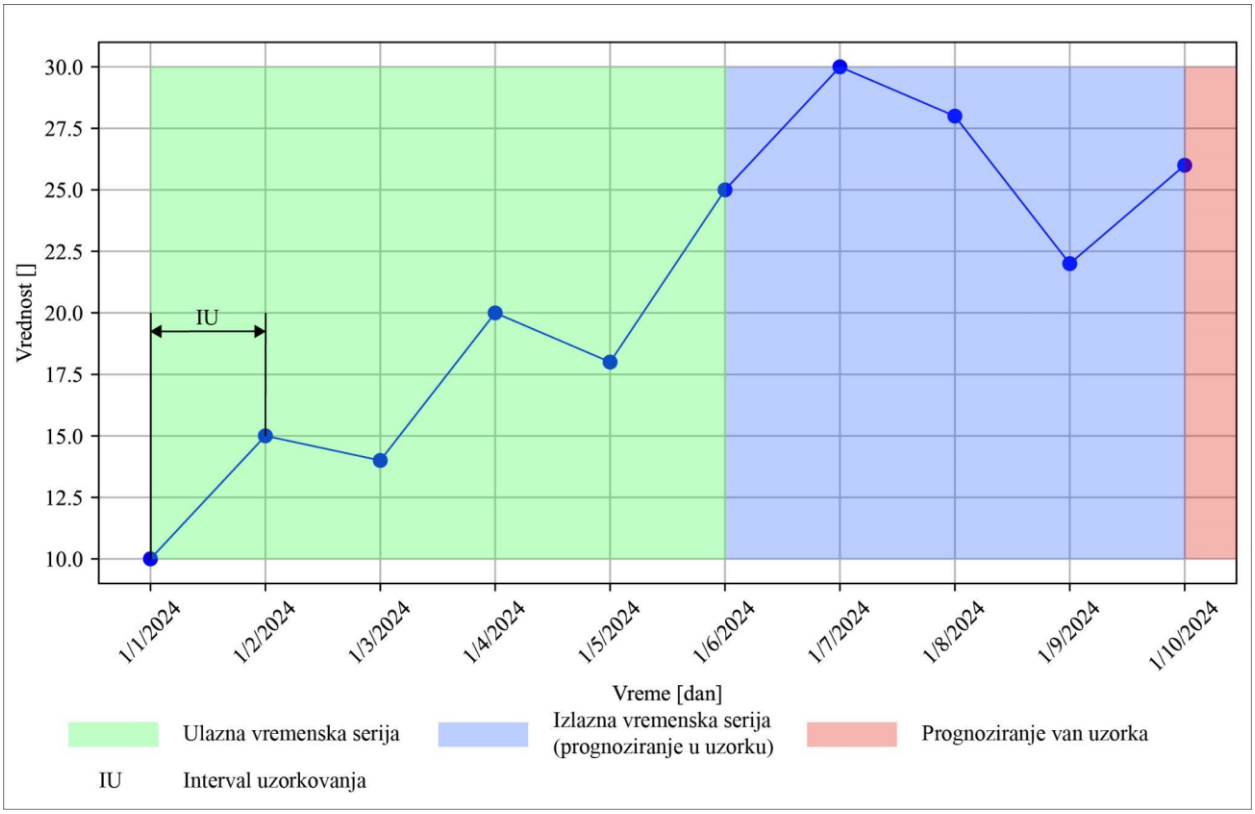
Metode prognoziranja vremenskih serija predstavljaju drugi vid modelovanja vođenim podacima. Za podatke se može reći da čine vremensku seriju ukoliko su prikupljeni više puta u određenom vremenskom periodu. Vremenski period može imati različite intervale, na primer visoku rezoluciju (desetine ili stotine delova sekunde) ili nisku rezoluciju (nedeljni, mesečni, kvartalni, godišnji itd.), u zavisnosti od problema koji se rešava. Bitan aspekt prognoziranja vremenskih serija jeste da su vremenski intervali između podataka jednaki[5], tj. da su podaci uzorkovani u redovnim vremenskim razmacima.

Cilj metoda prognoziranja vremenskih serija je predviđanje budućih vrednosti određenog parametra na osnovu njegovih istorijskih (prethodnih) vrednosti. Potrebno je napomenuti da metode prognoziranja vremenskih serija nisu u svojoj osnovi metode mašinskog učenja, iako se u poslednje vreme i metode mašinskog učenja koriste za prognozu vremenskih serija. U nastavku ovog poglavlja biće prikazani osnovni pojmovi, specifične vrste obrade, kao i Fejsbukov Profet (eng. *Facebook Prophet*) model koji je korišćen u ovoj disertaciji.

2.2.1. Osnovni pojmovi prognoziranja vremenskih serija

Metode prognoziranja vremenskih serija obično nemaju podelu na trening i test podatke, ali je za proveru datog modela moguće primeniti prognozu na uzorku, što omogućava ocenu kvaliteta modela. Nakon što se proceni da je prognoza zadovoljavajuća, može se izvršiti prognoza van uzorka. Slika 4[6] prikazuje slučaj vremenske serije koja je podeljena na ulaznu vremensku seriju (nalik trening setu podataka u mašinskom učenju) i izlaznu vremensku seriju (nalik test setu podataka u mašinskom učenju). Izlazna vremenska serija (ili, drugim rečima, prognozni horizont- eng. *Forecasting horizon*) koristi se za ocenu kvaliteta modela prognoziranja vremenskih serija. Za razliku od mašinskog učenja, metode prognoziranja vremenskih serija uglavnom nemaju atribute, već se prognoza zasniva samo na vremenskoj seriji.

[5] Moguće je prognozirati i nejednako uzorkovana vremenske podatke, ali to prevazilazi okvir ove disertacije. Uglavnom prilikom planiranja istraživanja, ukoliko se radi o vremenskim serijama istraživanje se projektuje tako da bude jednak interval uzorkovanja.
[6] Za sliku 4, u svrhe ilustracije, su korišćeni sintetički podaci.



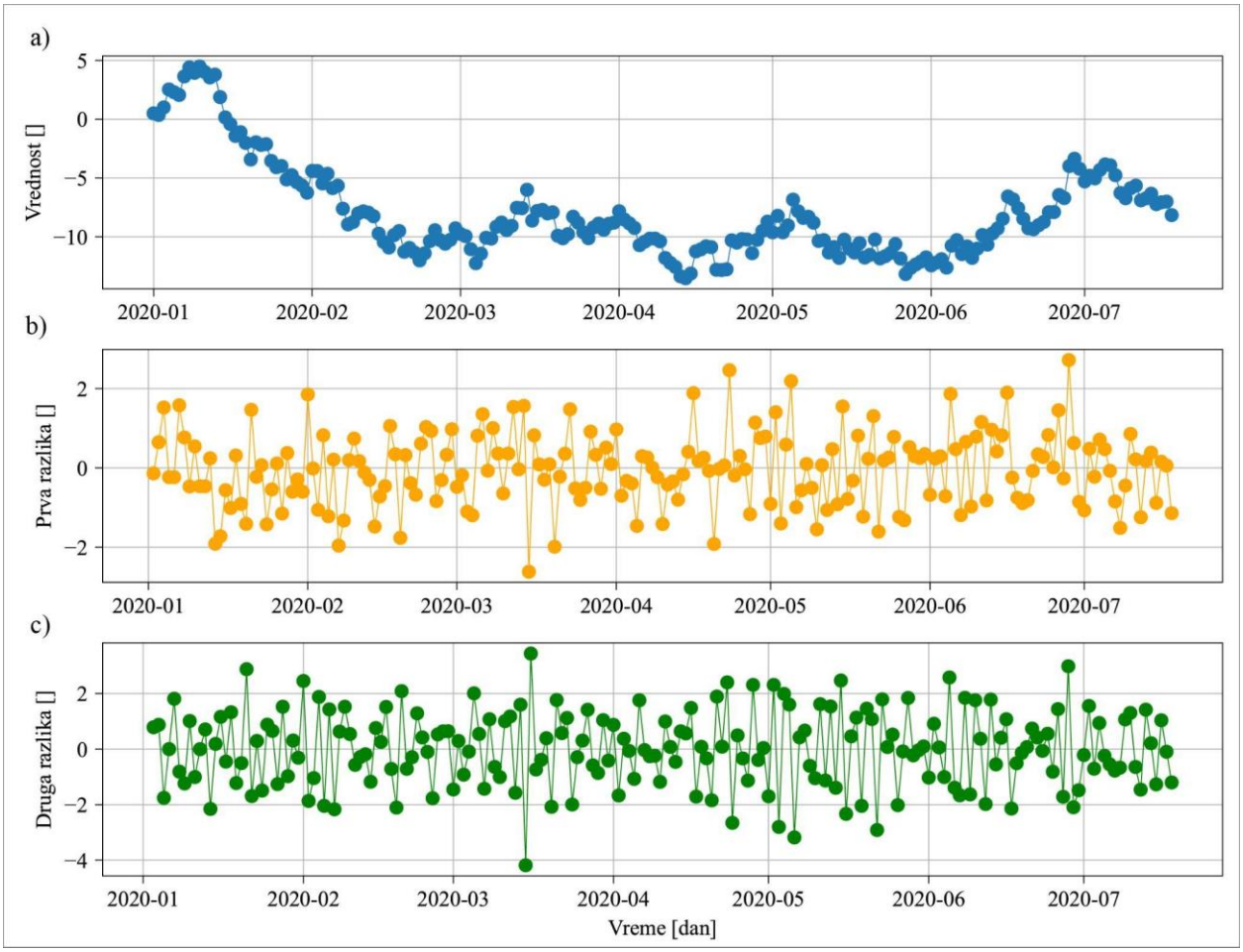
Slika 4. Pojednostavljeni prikaz elemenata vremenskih serija

Metode obrade podataka vremenskih serija nisu isključivo svojstvene za prognozu vremenskih serija, već su našle svoju primenu i u mašinskom učenju. Međutim, važno je napomenuti da su nastale upravo za potrebe metoda prognoziranja vremenskih serija. Jedna od često korišćenih metoda obrade podataka predstavlja uprošćavanje vremenske serije, odnosno dovođenje vremenske serije u stacionarnost. Stacionarnost vremenskih serija postiže se oduzimanjem svih komponenti vremenske serije (trend, cikličnost), tako da vremenska serija prikazuje samo oscilacije oko konstantne vrednosti. Tehnike za dobijanje stacionarne vremenske serije mogu biti jednostavne, poput logaritamske, kvadratne, kubne ili drugih korena vremenske serije, kao i prvi ili drugi diferencijal vremenske serije (što je najčešće primenjivana metoda). Za proveru stacionarnosti vremenske serije nakon transformacije moguće je primeniti prošireni Diki-Fulerov test (eng. *Augmented Dickey- Fuller test*) i/ili Kviatkovski-Filips-Šmit-Šin test (eng. *Kwiatkowski- Phillips- Smith- Shin* test). Oba testa, kao i prethodno prikazani Kolmogorov-Smirnov test, vrlo se lako mogu implementirati na podatke, a njihova interpretacija zasniva se na poređenju test statistike i kritične vrednosti za odabrani interval značajnosti. U praktičnim uslovima, ako je test statistika manja od kritične vrednosti za dati interval značajnosti, za vremensku seriju se može smatrati da je stacionarna. Razlika između proširenog Diki-

Fulerovog testa i Kviatkovski- Filips- Šmit- Šin testa ogleda se u nultoj hipotezi; nulta hipoteza za Diki- Fulerov test predstavlja nestacionarnost vremenske serije, dok za Kviatkovski- Filips- Šmit- Šin test predstavlja stacionarnost vremenske serije. Slika 5a[7] prikazuje vremensku seriju, dok Slike 5b i 5c prikazuju prvi i drugi diferencijal iste vremenske serije. Lako je primetiti da podaci sa slika 5b i 5c prikazuju relativno konstantne vrednosti sa malim fluktuacijama. Zbog ovog svojstva, modelovanje i prognoza stacionarnih vremenskih serija su lakši, što u teoriji daje bolje rezultate prognoza.

Drugi vid obrade podataka odnosi se na preskočene opservacije (eng. *Missing observations*). Preskočene opservacije predstavljaju merenja koja su trebala biti izmerena, ali su preskočena, ostavljajući prazninu u vremenskoj seriji. Preskočene opservacije smanjuju kvalitet podataka, pa je važno da istraživač adekvatno reši problem preskočenih opservacija. Postoje dve glavne opcije za rad sa preskočenim opservacijama: amputacija (uklanjanje podataka) ili imputacija (zamena preskočenih opservacija sa drugim vrednostima). Uglavnom, amputacija podataka se koristi kao poslednja opcija, na primer, u slučaju velikog broja preskočenih opservacija, dok se češće primenjuje imputacija podataka. Imputacija podataka predstavlja vrlo atraktivnu granu trenutnog istraživanja, sa velikim brojem metoda koje su razvijene i koje se trenutno razvijaju za ove potrebe. Jedna od najjednostavnijih metoda imputacije podataka je zamena preskočenih opservacija sa srednjom vrednošću ili medijanom preostalih podataka. Ovaj pristup ima svoje nedostatke, naročito kada se radi o imputaciji velikog broja podataka. Sa druge strane, ako je broj preskočenih opservacija mali, moguće je primeniti ranije pomenute metode ili druge, kao što su linearna interpolacija, Hermiteov interpolacioni polinom i druge. Takođe, treba napomenuti da preskočene opservacije mogu nastati i tokom čišćenja podataka (eng. *Data cleaning*), kada, na primer, opservacije nisu zaista preskočene, ali prikazuju šablon neadekvatnih merenja (npr. veoma niske ili visoke kontinualne vrednosti). Takve opservacije je potrebno ukloniti i zameniti drugim vrednostima.

[7] Za Sliku 5, u svrhe ilustracije, su korišćeni sintetički podaci.



Slika 5. (a) Prikaz vremenske serije; (b) Prvi diferencijal vremenske serije; (c) Drugi diferencijal vremenske serije

2.2.2. Fejsbuk Prophet model (eng. *Facebook Prophet Model*)

Fejsbukov Profet (eng. *Facebook Prophet*) model je razvijen od strane Fejsbuk platforme 2018. godine, sa ciljem prognoziranja parametara društvenih mreža koji su predstavljeni u formi vremenskih serija. Uvođenje ovog modela za prognozivanje vremenskih serija imalo je za cilj unapređenje opšteg modela prognoziranja vremenskih serija, koji može biti lako implementiran od strane radnika u privredi i istraživača koji imaju znanje o podacima koje koriste, ali nemaju duboko razumevanje metoda prognoziranja vremenskih serija (Taylor & Lentham, 2018). Jedan od ključnih ciljeva razvoja ovog modela bio je omogućiti generisanje velikog broja prognoza, kako bi istraživači imali efikasan i automatski način poređenja različitih prognoziranih vrednosti.

U svojoj osnovi, Fejsbukov Profet model se može predstaviti jednačinom:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \,, \tag{17}$$

Gde je $g(t)$ funkcija koja modeluje ne-periodične promene vremenske serije, $s(t)$ su periodične promene vremenske serije (nedeljne, mesečne itd.), $h(t)$predstavlja efekte praznika, odnosno nejednako raspoređene dane koji se razlikuju od drugih dana, a ϵ_t su promene u vremenskoj seriji koje imaju normalnu raspodelu, ali nisu modelovane od strane modela. U svojoj osnovi, Fejsbukov Profet model prilagođava teoretski sračunatu krivu prema krivi podataka, za razliku od klasičnih modela (npr. ARIMA modeli) koji preračunavaju vremenske odlike podataka (Taylor & Letham, 2018).

Pogodnosti Fejsbukovog Profet modela ogledaju se najpre u dve prethodno pomenute stavke koje prikazuju ograničenja klasičnih modela vremenskih serija- nejednako uzrokovani podaci i preskočene opservacije. Zbog suštinskih razlika u odnosu na klasične modele vremenskih serija, Fejsbukov Profet model ne zahteva jednako uzrokovane podatke, niti je potrebno da sve opservacije budu popunjene[8]. Zbog ovih karakteristika, Fejsbukov Profet model je našao široku primenu u nauci i privredi, posebno u prognoziranju različitih vremenskih serija. Za prognoziranje vremenskih serija, bitno je i oceniti model, odnosno, kada se prognoza vrši unutar uzorka, prognozirane vrednosti treba uporediti sa stvarnim vrednostima pre nego što se prognozira van uzorka. Pošto su metode za evaluaciju modela regresije iste kao i metode za evaluaciju modela vremenskih serija (oba predstavljaju slučaje regresije), u ovom poglavlju se neće ponavljati iste metode.

Slično kao i za metode mašinskog učenja, metode prognoziranja vremenskih serija spadaju u nauku o podacima, primenjenu statistiku i druge metode koje nose različita imena, ali imaju isti cilj. U poglavlju metodologije ove disertacije prikazana je generalna ideja i ciljevi prilikom modelovanja vođenim podacima, ali sveobuhvatan prikaz ove grane nauke o podacima nije prikazan, jer to izlazi van okvira ove disertacije. Na primer, u okviru metoda prognoziranja vremenskih serija postoji veliki broj drugih modela kao što su ARIMA, ARCH i GARCH

[8] Potrebno je napomenuti da je kvalitet podataka presudan za modele koji su vođeni podacima. Prisustvo što većeg procenta podataka za dati vremenski interval kao i jednako uzorkovani podaci predstavljaju teoretsko olakšanje modelu da modeluje podatke, a ujedno time i preciznije prognoze i smanjene greške.

modeli, Holt- Vintersova metoda i drugi. Takođe, postoji veliki broj načina za dobijanje stacionarne vremenske serije, pored prikazanih jednostavnih metoda, kao što su Boks-Koks metoda, dekompozicija vremenskih serija i drugi. Pored toga, nije bilo previše reči o kvalitetu podataka, uklanjanju opservacija koje se smatraju neadekvatnim, kao i metodama označavanja velikog broja podataka. Sve prethodno pomenute teme predstavljaju vrlo bitne aspekte u okviru nauke o podacima, ali nisu eksplicitno tema ove disertacije. Takođe, prilikom primene nauke o podacima za neku drugu, specifičnu nauku, kao što su geonauke, geofizika i atmosferska fizika i druge srodne naučne grane, potrebno je imati znanje o prirodi podataka, kao i o metodama nauke o podacima, odnosno kako i kada primeniti određene metode.

3. Rezultati

Rezultati sprovedenih istraživanja prikazani su u poglavlju 3. Rezultati su grupisani i razvrstani prema oblastima istraživanja kojima pripadaju. Prvo podpoglavlje (3.1.) odnosi se na rezultate primene modela vođenih podacima na podatke koncentracije zagađujućih materija u vazduhu, drugo podpoglavlje (3.2.) na prostornu klasifikaciju ofiolita istočne Vardarske zone, treće podpoglavlje (3.3.) na primenu modela vođenih podacima u fizici jonosfere, dok poslednje podpoglavlje (3.4.) predstavlja primenu statističkih metoda na podatke magnetne susceptibilnosti sa jalovišta rudnika „Rudnik“.

3.1. Primena modela vođenih podacima na podatke koncentracije zagađujućih materija u vazduhu u Republici Srbiji

Termin *kvalitet vazduha* obuhvata širok spektar parametara koji kvantifikuju različite frakcije komponenata prisutnih u vazduhu. U užem smislu, kvalitet vazduha se može predstaviti kroz koncentraciju zagađujućih materija, kao što su PM_{10} (čestice do 10 μm) i $PM_{2.5}$ (čestice do 2,5 μm). Značajnost ovih parametara ogleda se u veličini čestica koje su suspendovane u vazduhu: manje čestice imaju veći potencijal da prodiru dublje u respiratorni sistem (Rahman et al., 2023; Zhang et al., 2023), čime nanose veću štetu ljudima koji udišu takav zagađeni vazduh. Vazduh sa visokim nivoom zagađujućih materija predstavlja značajan rizik za celokupno stanovništvo, jer su povećane vrednosti ovih materija povezane sa većim mortalitetom (Dockery et al., 1992; Araujo, 2011), kardiovaskularnim bolestima (Araujo, 2011), izraženijom susceptibilnošću na alergene (Bernstein, 2004), kao i bolestima respiratornog sistema (Libasin et al., 2020; Rakholia et al., 2022). Razlozi za povećane vrednosti zagađujućih materija u vazduhu su različiti. Sa jedne strane, industrijalizacija, urbanizacija i rast stanovništva doprinose povećanju zagađenja (Harishkumar et al., 2020; Wardana et al., 2022; Zhang & Zhang, 2023), dok sa druge strane, sagorevanje fosilnih goriva, neregulisana energetska industrija i upotreba neadekvatnih materijala za grejanje domaćinstava dodatno povećavaju nivo zagađujućih materija u vazduhu u Republici Srbiji.

U Republici Srbiji, najveći broj stanica za automatski kontinualni monitoring parametara koncentracije zagađujućih materija u vazduhu obavlja Agencija za zaštitu životne sredine (SEPA). U trenutku pisanja doktorske disertacije, SEPA ima ukupno 76 stanica za monitoring

koncentracije zagađujućih materija u vazduhu, od kojih se 34 nalaze u Beogradu. U nastavku ovog poglavlja biće prikazana dva primera vezana za koncentraciju zagađujućih materija u vazduhu. Prvi primer se odnosi na primenu Fejsbukovog *Prophet* modela za prognoziranje budućih vrednosti koncentracije zagađujućih materija u vazduhu, dok drugi primer prikazuje primenu metoda mašinskog učenja za imputaciju preskočenih opservacija koncentracije zagađujućih materija u vazduhu.

Prvi primer je od velike važnosti za monitoring koncentracije zagađujućih materija u vazduhu, jer spoznaja o prostornoj i vremenskoj komponenti budućih vrednosti koncentracije zagađujućih materija u vazduhu predstavlja ključnu informaciju za ugrožene grupe, kao i za regulativna tela i donosioce odluka u Republici Srbiji. Sa druge strane, automatski kontinualni monitoring koncentracije zagađujućih materija u vazduhu često je opterećen smetnjama u merenju, koje se najčešće ogledaju u prekidima mernog signala i preskočenim opservacijama. Zbog toga će, ovo poglavlje obuhvatiti dva primera: prvi se odnosi na prognoziranje budućih vrednosti koncentracije zagađujućih materija u vazduhu (3.1.1.), a drugi na razvoj metode za poboljšanje kvaliteta podataka koncentracije zagađujućih materija u vazduhu (3.2.2.).

3.1.1. Primena Fejsbukovog Profet modela za prognoziranje budućih vrednosti koncentracije zagađujućih materija u vazduhu na mernoj stanici Beograd-Zeleno brdo

Fejsbukov *Prophet* model (opisan u 2.2.2.) se pokazao veoma korisnim za prognoziranje vrednosti koncentracije zagađujućih materija u vazduhu (Ye, 2019; Samal et al., 2019; Zhoul et al., 2020; Tejasvini et al., 2020; Shen et al., 2020). Neki od tih rezultata prikazuju primenu Profet modela u kombinaciji sa drugim modelima, čime se dobija hibridni model koji u određenim slučajevima daje veću tačnost nego sam Profet model. Cilj istraživanja je primena Profet modela na podatke o koncentraciji zagađujućih materija u vazduhu (koncentracije $PM_{2.5}$ i PM_{10}) koji su prikupljeni u Republici Srbiji od strane Agencije za zaštitu životne sredine (SEPA) za mernu stanicu Beograd-Zeleno brdo.

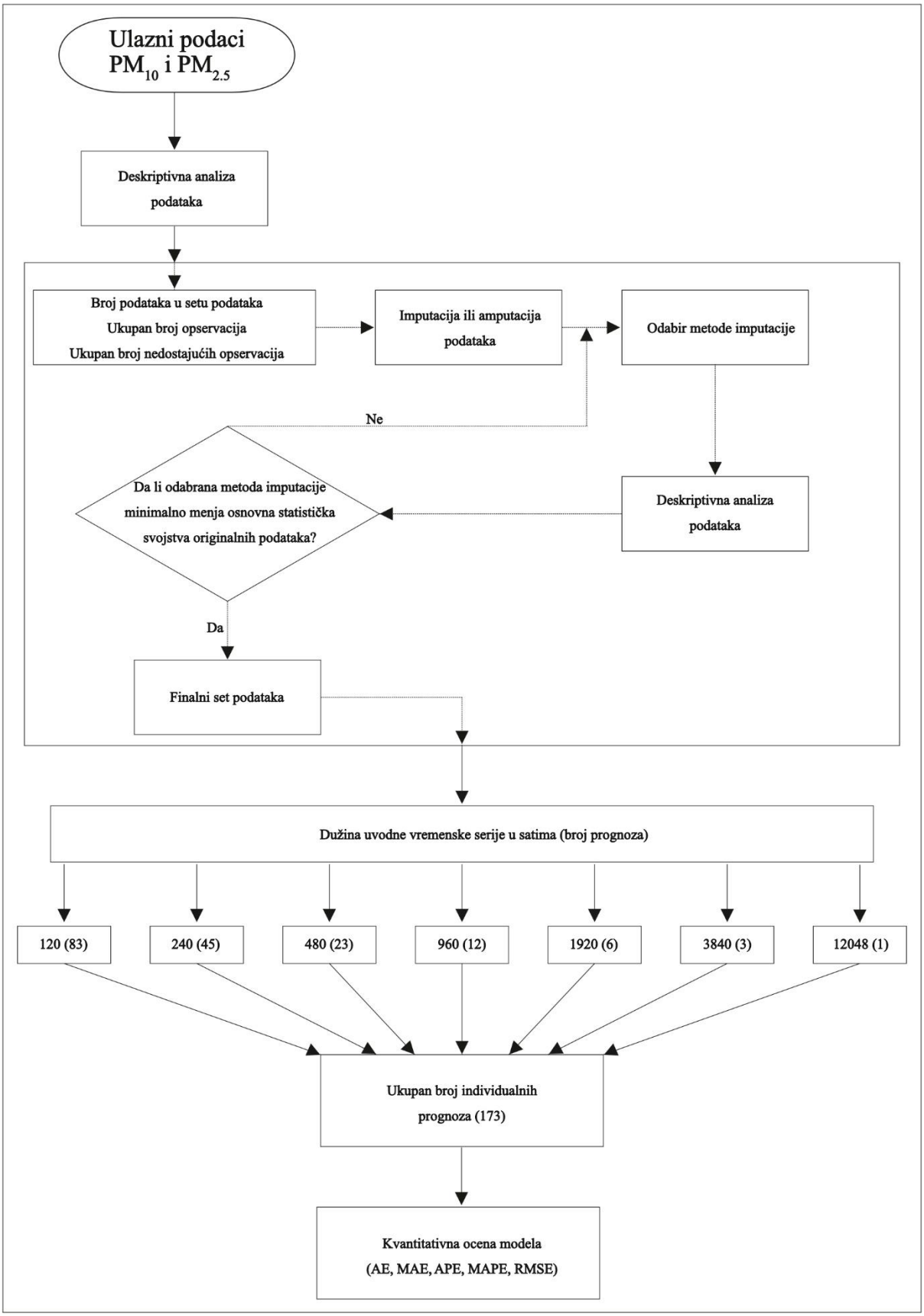
3.1.1.1. Postavka i radni tok istraživanja

Ulazni podaci satnih vrednosti koncentracija PM_{10} i $PM_{2.5}$ čestica za lokaciju Beograd-Zeleno brdo dobijeni su od strane Agencije za zaštitu životne sredine za period od 1. januara 2021. godine do 17. avgusta 2022. godine. Prvi korak u analizi podataka predstavlja deskriptivna

statistika, odnosno utvrđivanje srednje vrednosti, medijane, minimuma, maksimuma i drugih osnovnih parametara deskriptivne statistike (Slika 6).

Sa obzirom na to da je poznato da podaci o kvalitetu vazduha mogu sadržati preskočene opservacije, pored prethodno pomenutih parametara, bilo je potrebno utvrditi i broj preskočenih opservacija u skupu podataka, što je ključno za donošenje odluke o imputaciji i/ili amputaciji podataka. Ukoliko se za određeni skup podataka odluči za imputaciju, istraživanje je postavljeno tako da metoda imputacije ne menja parametre deskriptivne statistike nakon što se vrednosti pridodaju preskočenim opservacijama.

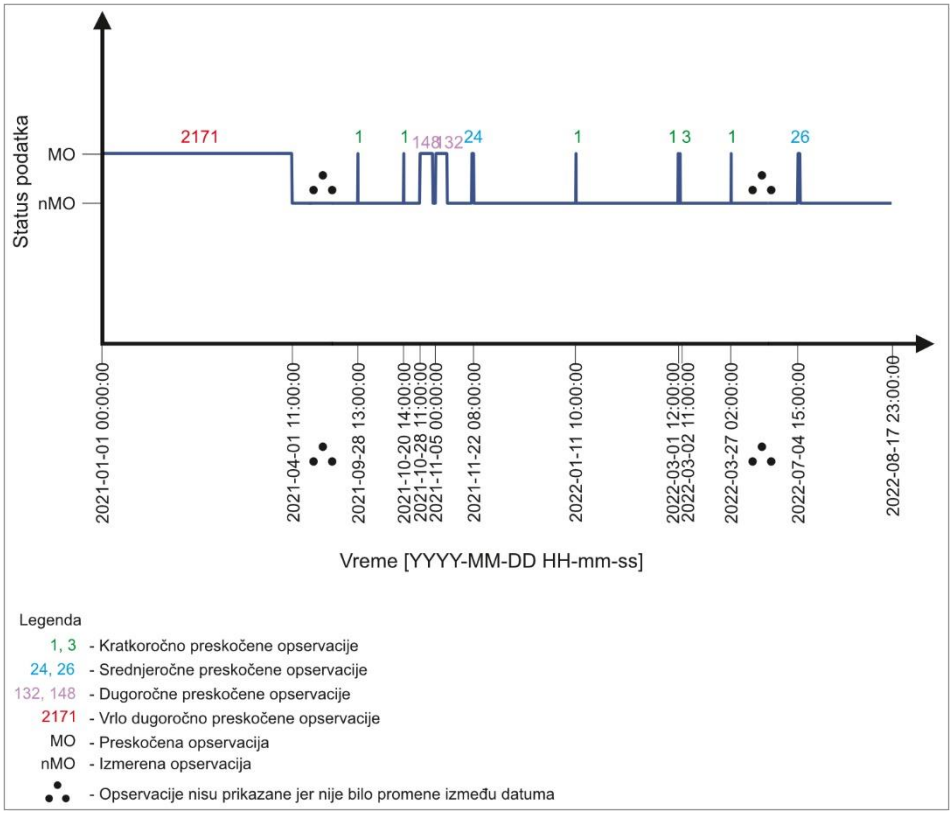
Nakon dobijanja finalnog seta podataka, prognoziranje je vršeno u više etapa. Prva etapa podrazumevala je korišćenje prvih 120 sati kao uvodne vremenske serije za prognoziranje narednih 24 sata. Nakon prve prognoze, uzimaju se narednih 120 podataka, i proces se ponavlja. U prvoj etapi generisano je ukupno 83 prognoze. Druga etapa podrazumevala je povećanje dužine uvodne vremenske serije sa 120 na 240 sati, čime je generisano ukupno 45 prognoza. Proces prognoziranja u etapama sproveden je dok nije iskorišćen ceo set podataka (12048 podataka) za prognozu narednih 24 sata. Ukupan broj individualnih prognoza iznosio je 173, što je omogućilo detaljnu statističku analizu odstupanja prognoza Profet modela od stvarnih vrednosti. Za ocenu kvaliteta modela korišćene su metode za kvantifikaciju koje su prethodno prikazane u poglavlju 2.1.2.12.



Slika 6. Radni tok istraživanja prognoziranja koncentracije zagađujućih materija u vazduhu na stanici Beograd- Zeleno brdo korišćenjem Fejsbukovog Profet modela

3.1.1.2. Obrada i istraživačka analiza podataka

Faza obrade podataka i istraživačke analize podataka prvo je korišćena za formiranje finalnog skupa podataka koji ne sadrži preskočene opservacije. Slika 7 prikazuje raspodelu preskočenih opservacija i izmerenih podataka, gde se može videti da podaci od 1. januara 2021. godine do 1. aprila 2021. godine nisu izmereni. Zbog toga je odlučeno da se ti podaci amputiraju, tj. popunjavanje 2171 uzastopnog sata preskočenih opservacija nije bilo moguće izvesti. Ostale preskočene opservacije kategorizovane su u tri grupe: kratkoročne preskočene opservacije u trajanju do 3 uzastopna sata, srednjeročne preskočene opservacije u trajanju od 24 do 26 sati i dugoročne preskočene opservacije u trajanju od 132 i 148 uzastopnih sati. Ukoliko se izuzme grupa od 2171 uzastopnog preskočenog sata, u setu podataka postoji ukupno 338 preskočenih opservacija, što čini samo 2,8% od celokupnog skupa podataka.



Slika 7. Raspodela preskočenih opservacija u istražnom periodu

Za kratkoročne preskočene opservacije korišćena je linearna interpolacija, jer je prikazala rezultate koji manje menjaju parametre deskriptivne statistike u poređenju sa metodom „*poslednja opservacija preneta unapred*“. Srednjoročne preskočene opservacije popunjene su

primenom imputacije srednjom vrednošću, pri čemu su uzimane satne srednje vrednosti, što je omogućilo da preskočene opservacije u trajanju od 24 i 26 sati imaju određene dnevne varijacije. Promene parametara deskriptivne statistike ogledaju se u medijalnoj vrednosti za PM_{10} parametar, kao i u koeficijentu spljoštenosti i asimetrije za oba parametra koncentracije zagađujućih materija u vazduhu, ali te promene ne prelaze 0,7% (Tabela 2). Za dugoročne preskočene opservacije korišćen je Hermiteov interpolacioni polinom, koji je, od svih primenjenih metoda, pokazao minimalno odstupanje parametara deskriptivne statistike u odnosu na one izračunate na skupu podataka bez popunjenih preskočenih opservacija.

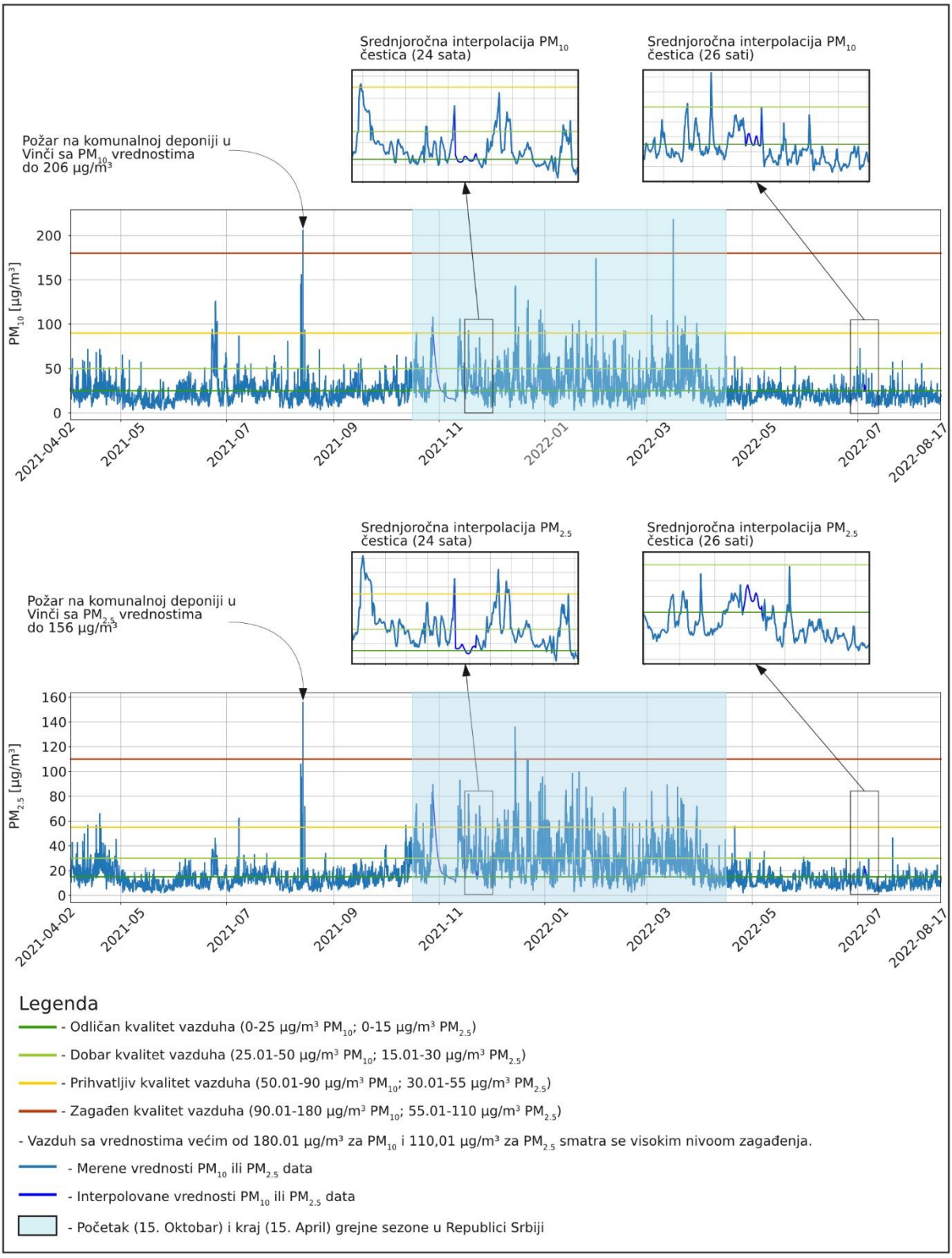
Tabela 2. Deskriptivna statistika sračunata pre i nakon imputacije podataka

Metoda	**Ulazni podatak**		**Linearna interpolacija**		**Satni prosek**		**Hermiteov polinom**	
Dužina preskočenih opservacija	/		**Kratkoročno**		**Srednjoročno**		**Dugoročno**	
Parametar	**PM_{10}**	**$PM_{2.5}$**	**PM_{10}**	**$PM_{2.5}$**	**PM_{10}**	**$PM_{2.5}$**	**PM_{10}**	**$PM_{2.5}$**
Broj podataka [/]	11746	11746	11754	11754	11804	11804	12084	12084
Minimum [$\mu g/m^3$]	2,97	1,94	2,97	1,94	2.97	1.94	2.97	1.94
Maksimum [$\mu g/m^3$]	218	156	218	156	218	156	218	156
Srednja vrednost [$\mu g/m^3$]	26,44	18,69	26,44	18,69	26.44	18.69	26.53	18.85
Medijana [$\mu g/m^3$]	22,40	14,60	22,40	14,60	22.50	14.60	22.50	14.70
Modalitet [$\mu g/m^3$]	21,50	11,00	21,50	11,00	21.50	11.00	21.50	11.00
Koeficijent asimetrije [/]	2,39	2,07	2,39	2,07	2.40	2.08	2.37	2.07
Koeficijent spljoštenosti [/]	11,20	6,30	11,20	6,30	11.26	6.34	10.82	6.13
Test modaliteta [/]	T		T		T		T	
Pirsonov KK [/]	0,87		0,87		0,87		0,87	
Spirmanov KK [/]	0,87		0,87		0,87		0,87	

Prilikom primene drugih metoda imputacije podataka za dugoročne preskočene opservacije, kao što su K- najbližih suseda ili imputacija srednjom vrednošću, raspodela podataka prelazila je iz unimodalne u bimodalnu, čime su značajno promenjeni parametri deskriptivne statistike i sama raspodela podataka. Sa druge strane, primenom Hermiteovog interpolacionog polinoma za imputaciju dugoročnih preskočenih opservacija, raspodela je ostala unimodalna (Tabela 2). Takođe, još jedan vid provere može se pronaći u konstantnoj vrednosti koeficijenata korelacije, koji, pored toga što ukazuju na jaku korelaciju između PM_{10} i $PM_{2.5}$ čestica, takođe ukazuju na linearan odnos između ta dva parametra.

Slika 8 prikazuje finalni set podataka nakon imputacije. U periodu od 15. oktobra 2021. godine do 15. aprila 2022. godine, vizuelno se može uočiti najveće povećane vrednosti $PM_{2.5}$ čestica,

a potom u manjoj meri, i PM_{10} čestica. Analiza korelacije temperature za isti vremenski period i rezoluciju, koja je dobijena sa Beogradskog aerodroma „Nikola Tesla“, daje vrednosti korelacije za PM_{10} čestice u iznosu od -0,25 (Pirson) i -0,26 (Spirman), dok se za $PM_{2.5}$ čestice dobijaju vrednosti od -0,52 (Pirson) i -0,56 (Spirman). Prema klasifikaciji koeficijenta korelacije prema Schober et al. (2018), za korelaciju PM_{10} čestica i temperature može se reći da je slaba i negativna, dok za $PM_{2.5}$ čestice ta korelacija predstavlja srednju i negativnu vrednost. Drugim rečima, sa smanjenjem temperature tokom godine, dolazi do slabog povećanja koncentracija PM_{10} čestica i srednjeg povećanja $PM_{2.5}$ čestica, verovatno usled korišćenja ogrevnih materijala u domaćinstvima.



Slika 8. Konačni skup PM_{10} i $PM_{2.5}$ podataka korišćen za istraživanje sa prikazanim odabranim karakteristikama signala

3.1.1.3. Rezultati prognoziranja vrednosti PM_{10} i $PM_{2.5}$ čestica

Tabela 3 prikazuje rezultate prognoziranja, gde su prikazani svi odnosi ulazne vremenske serije, kao i određene statistike za koren srednje kvadratnog odstupanja. Profet model u određenim slučajevima pokazuje vrlo neadekvatne vrednosti prognoze, kao što je slučaj za PM_{10} čestice sa dužinom ulazne vremenske serije od 480 podataka, gde je maksimalna vrednost korena srednjeg kvadratnog odstupanja iznosila 50,78 μg/m³. U toj situaciji, srednja apsolutna greška iznosila je 49,54 μg/m³, dok je procentualna greška bila 57,34%. Sa druge strane, u tom validacionom prozoru zabeležene su neočekivano velike vrednosti PM_{10} parametra, sa srednjom vrednošću od 85 μg/m³, dok je za ceo set podataka srednja vrednost PM_{10} čestica bila 26,53 μg/m³.

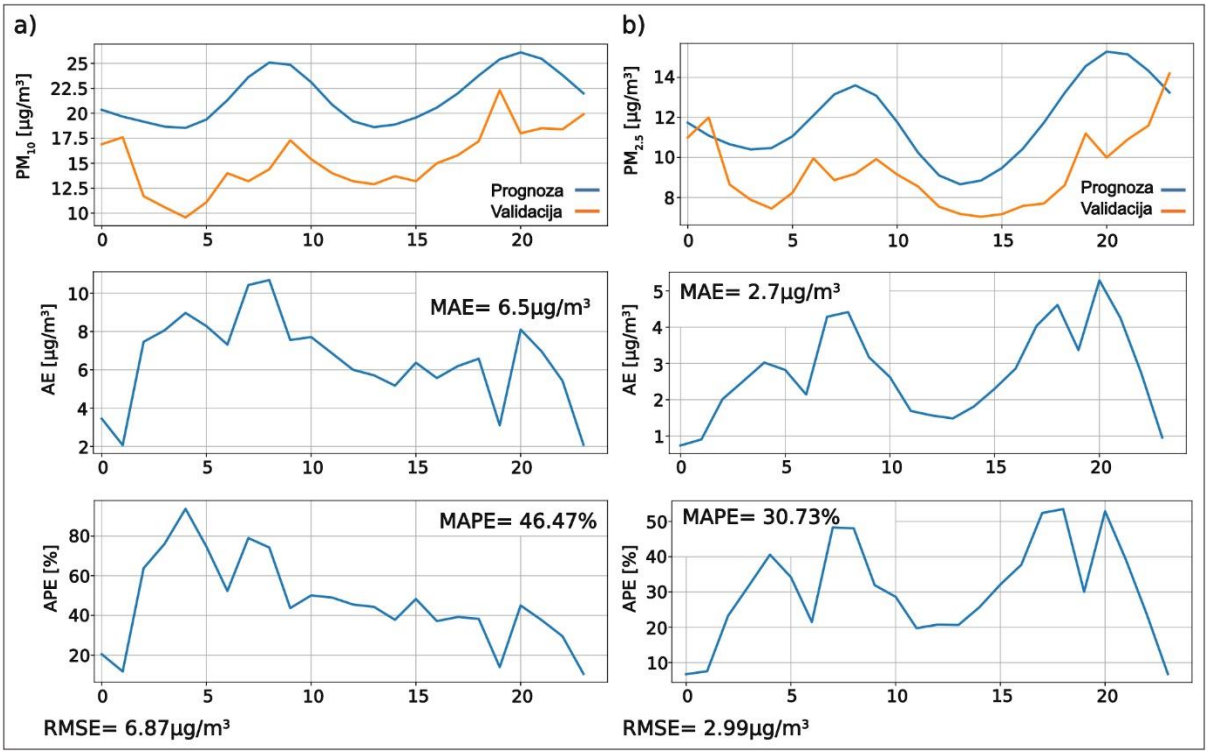
Tabela 3. Rezultati prognoziranja u vidu deskriptivne statistike raspodele korena srednjeg kvadratnog odstupanja

Parametar	Odnos	Broj prognoza	$RMSE_{min}$ [μg/m³]	$RMSE_{sred}$ [μg/m³]	$RMSE_{med}$ [μg/m³]	$RMSE_{max}$ [μg/m³]
$PM_{2.5}$	120:24 (5:1)	83	1,04	9,20	6,35	33,81
	240:24 (10:1)	45	1,41	8,45	5,82	38,71
	480:24 (20:1)	23	2,85	11,42	9,04	46,02
	960: 24 (40:1)	12	1,96	10,49	8,29	31,03
	1920:24 (80:1)	6	1,98	4,83	4,87	7,65
	3840:24 (160:1)	3	2,00	6,24	3,78	12,93
PM_{10}	120:24 (5:1)	83	1,64	12,85	10,79	41,21
	240:24 (10:1)	45	2,30	11,01	8,21	46,11
	480:24 (20:1)	23	3,77	16,19	15,26	50,78
	960: 24 (40:1)	12	3,77	15,23	11,52	49,25
	1920:24 (80:1)	6	3,12	6,47	6,17	9,39
	3840:24 (160:1)	3	4,49	8,97	7,04	15,39

Medijalna vrednost korena srednjeg kvadratnog odstupanja za $PM_{2.5}$ čestice pokazuje da je koren srednje kvadratnog odstupanja za sve odnose ulazne vremenske serije manje od 10 μg/m³ (drugim rečima, model u 50% prognoza, prognozira vrednost ispod 10 μg/m³). U slučaju odnosa ulazne vremenske serije 80:1, mogu se videti i najmanje vrednosti maksimalnog korena srednjeg kvadratnog odstupanja, koje iznose svega 7,65 μg/m³. Zbog relativno malog broja prognoza (samo 6), nije moguće sa velikom pouzdanošću tvrditi da je model postigao ovako male vrednosti odstupanja zbog dobijanja idealnog odnosa, a ne nasumično. Zbog toga je

potrebno sprovesti dodatna istraživanja sa većim brojem podataka i stanica kako bi se dobila detaljnija raspodela greške modela.

Slika 9 prikazuje vrednosti prognoza za ceo set podataka, pri čemu je ulazna vremenska serija imala dužinu od 12084 sata, dok je izlaz bio konstantnih 24 sata. Vrednosti apsolutne greške za PM_{10} parametar variraju između 2 i 10 μg/m³, sa prosečnom vrednošću od 6,5 μg/m³, dok vrednosti relativne greške variraju između 10,5% i 93%, sa prosečnom vrednošću od 46,45%. Vrednosti apsolutne greške za $PM_{2.5}$ variraju između 0,73 μg/m³ i 5 μg/m³, sa srednjom vrednošću od 2,7 μg/m³. Primer sa Slike 9 predstavlja dobar pokazatelj generalne neadekvatnosti korišćenja samo apsolutne i relativne greške za kvantifikaciju odstupanja prognozirane i stvarne vrednosti koncentracije zagađujućih materija u vazduhu. Na primer, na slici 9b, petnaesti sat ima stvarnu vrednost od 7,04 μg/m³, dok je prognoza bila 8,84 μg/m³. Apsolutno odstupanje nije preveliko (1,8 μg/m³), dok je relativna greška 25%. Za kvantifikaciju odstupanja prognozirane i stvarne vrednosti koncentracije zagađujućih materija u vazduhu potreban je parametar kao što je koren srednje kvadratnog odstupanja, koje ne zavisi u tolikoj meri od reda veličine stvarne i prognozirane vrednosti.



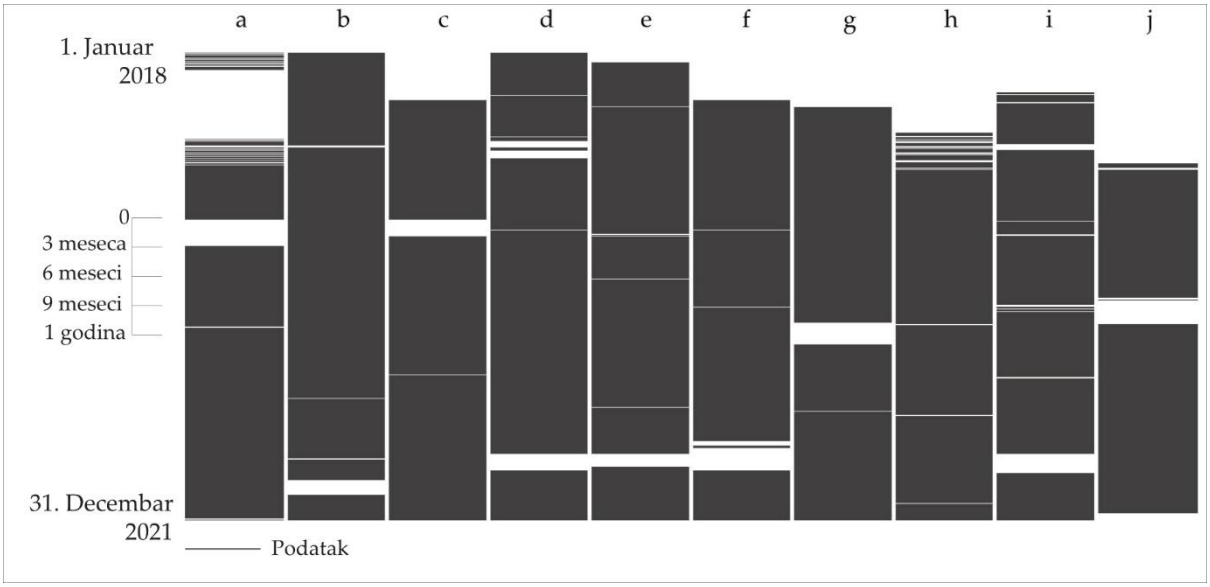
Slika 9. Prognozirana i istinita vrednost koncentracije zagađujućih materija u vazduhu za 24 sata nakon najduže uvodne vremenske serije

Sa druge strane, pored kvantitativnih mera odstupanja, vrednosti koncentracije zagađujućih materija u vazduhu mogu se razmatrati i kroz kvalitativne pokazatelje, kao što je vizuelna identifikacija varijacije prognoze modela. U oba primera sa Slike 9 može se videti da prognoza modela prikazuje određenu dnevnu varijaciju, koja odražava uopštenu sliku varijacije koncentracije zagađujuće materije u vazduhu tokom dana.

3.1.2. Primena metode slučajne šume za dvosmernu imputaciju podataka koncentracije zagađujućih materija u vazduhu u Republici Srbiji

U prethodnom primeru prikazano je na podacima jedne merne stanice kako istraživači nailaze na problem preskočenih opservacija prilikom rukovanja sa podacima o koncentraciji zagađujućih materija u vazduhu. U ovom primeru, takav problem može biti prikazan šire (Slika 10), gde se na primeru 10 stanica u četvorogodišnjem periodu vidi koliko zapravo predstavljaju preskočene opservacije. Na primeru sa Slike 10, u proseku stanice za merenje koncentracije zagađujućih materija u vazduhu sadrže oko 15,3% preskočenih opservacija, pri čemu je minimalna vrednost 3,9% (Beograd- Stari grad), dok je maksimalna 30,9% (Niš, osnovna škola „Sveti Sava“).

Zbog problema koje preskočene opservacije predstavljaju istraživačima, razvijanje metoda za imputaciju podataka o koncentraciji zagađujućih materija u vazduhu je od velikog značaja. Različite metode su razvijene kako bi se dobile pouzdane procene preskočenih opservacija (Junninen et al., 2004; Norazian et al., 2008; Jiang et al., 2020; Wijesekara & Liyange, 2020; Kim et al., 2021; Alsaber et al., 2021; Belaschsen et al., 2022; Chen et al., 2022; Flores et al., 2023; Kebalepile et al., 2024 i dr.), ali nijedna metoda do sada nije u potpunosti tačna, precizna i pouzdana. Jednostavne metode, poput imputacije podataka srednjom vrednošću ili medijanom celokupnog dostupnog seta podataka, mogu dovesti do vrlo loših procena preskočenih opservacija. Glavni problem jednostavnih metoda je to što ne uzimaju u obzir dnevne, nedeljne i mesečne varijacije koncentracije zagađujućih materija u vazduhu, koje mogu nastati prilikom popunjavanja većih količina preskočenih opservacija. Sa druge strane, pri popunjavanju manjeg broja preskočenih opservacija (kao što je bio slučaj u prethodnom primeru), moguće je koristiti jednostavne metode kako bi se postigla efikasnost i smanjilo računarsko vreme.



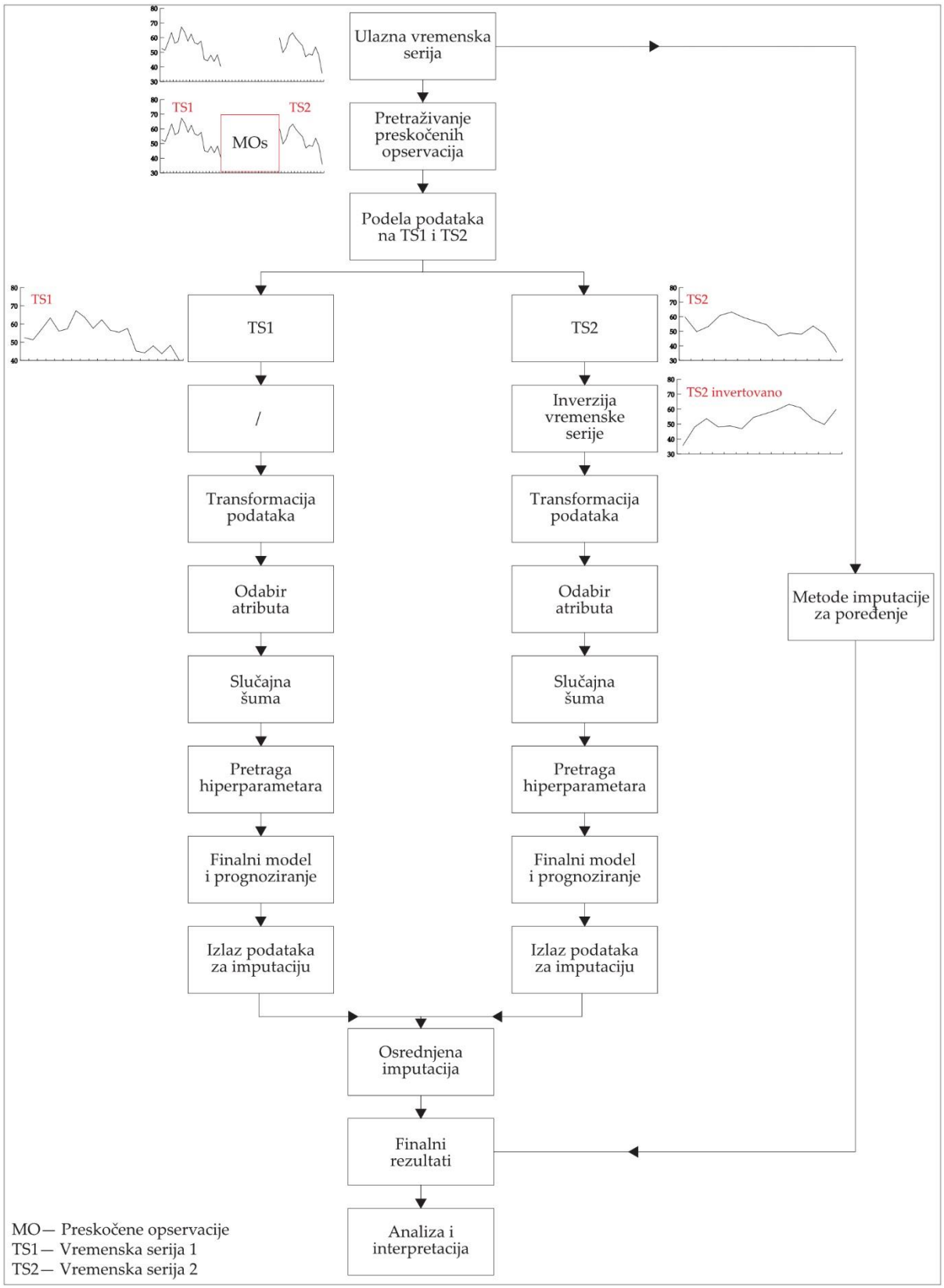
Slika 10. Preskočene opservacije $PM_{2.5}$ na odabranim stanicama za merenje koncentracije zagađujućih materija u vazduhu od 1. januara 2018. do 31. decembra 2021. godine; a- Novi Sad Rumenačka ulica; b- Beograd Stari Grad; c- Beograd Novi Beograd; d- Beograd Mostar; e- Smederevo Centar; f- Obrenovac Centar; g- Valjevo; h- Bor Gradski park; i- Kosjerić; j- Niš Osnovna škola „Sveti Sava“

U ovom slučaju cilj je bio pokušaj razvijanja metode za imputaciju koncentracija zagađujućih materija u vazduhu primenom metoda mašinskog učenja. Glavna ideja ovog pristupa sastoji se u tome da se problem imputacije podataka tretira kao problem prognoze jedne promenljive bez dodatnih atributa, pri čemu se prognoza vrši u dva smera. Prvi smer predstavlja klasično prognoziranje u „budućnost“, dok drugi smer podrazumeva prognozu u „prošlost“. Ideja za razvoj ovakve metode proističe iz toga što dodatni podaci nisu uvek dostupni, pa je metoda koja se zasniva samo na datoj promenljivoj vrlo značajna. Sa druge strane, podaci koji se nalaze nakon preskočenih opservacija uglavnom se ne koriste, ali u ovom slučaju oni služe kao dodatna prognoza.

3.1.2.1. Radni tok algoritma, podaci i postavka istraživanja

Radni tok algoritma prikazan je na slici 11. Ulazni parametar je vremenska serija koja sadrži određene preskočene opservacije. Algoritam najpre pretražuje vremensku seriju kako bi pronašao lokaciju i dužinu preskočenih opservacija. Nakon pronalaska lokacije i dužine preskočenih opservacija, algoritam deli celokupnu vremensku seriju na dva dela: TS1 i TS2. Obrada vremenskih serija TS1 i TS2 se razlikuje u jednom koraku, u kojem se vrši inverzija vremenske serije TS2. Inverzija vremenske serije vrši se jednostavno, tako što se njen redosled

obrće kako bi se omogućilo prognoziranje u budućnost ili, u ovom slučaju, u prošlost. Transformacija podataka primenjuje se na obe vremenske serije, a ogleda se u logaritamskoj transformaciji podataka.



Slika 11. Radni tok predloženog algoritma i postavka istraživanja

Odabir atributa, kao što je prethodno prikazano, predstavlja jedan od ključnih koraka u modelovanju podataka mašinskim učenjem. U ovom slučaju, pošto su postavljeni uslovi za istraživanje, koji se ogledaju u neupotrebi dodatnih atributa poput meteoroloških uslova, parametara saobraćaja i drugih parametara koncentracije zagađujućih materija u vazduhu, jedini atributi koji se mogu koristiti su atributi vezani za same vrednosti $PM_{2.5}$ čestica. Zbog toga, dostupni atributi obuhvataju prethodne vrednosti signala, klizajuće statistike (kliznu srednju vrednost, medijanu i standardnu devijaciju) i druge.

Nakon odabira atributa i dobijanja finalnog skupa podataka, podaci su prosleđeni na model slučajne šume koji koristi 80% podataka za treniranje modela, a preostalih 20% za testiranje i dobijanje vrednosti za imputaciju. Pronalazak hiperparametara vršen je nasumičnim pretraživanjem, pri čemu je odabran broj stabala u opsegu od 10 do 1000. Prilikom generisanja novih vrednosti, algoritam ih dodaje u listu podataka i atributa kako bi iterativno generisao narednu vrednost, sve dok broj generisanih vrednosti ne bude izjednačen sa brojem preskočenih opservacija u originalnoj vremenskoj seriji. Nakon dobijanja dva izlaza podataka za imputaciju- jedan koji odgovara vremenskoj seriji TS1, a drugi vremenskoj seriji TS2- oba izlaza se osrednjavaju kako bi se dobio treći izlaz iz algoritma.

Finalni rezultati, analiza i interpretacija dobijaju se nakon primene metoda imputacije za poređenje. Istraživanje bez metoda za poređenje ne bi prikazalo značajne rezultate pogodne za interpretaciju. Zbog toga je u ovom istraživanju odlučeno da se porede dva jednostavna pristupa: imputacija podataka srednjom vrednošću i medijalnom vrednošću. Razlog za korišćenje ovih dvaju jednostavnih metoda leži u dva aspekta. Ukoliko razvijeni algoritam ne može da pruži bolje rezultate od jednostavnih modela, poput imputacije srednjom vrednošću i medijalnom vrednošću, tada algoritam nije vredan daljeg korišćenja. Sa druge strane, ukoliko algoritam pokaže približno iste ili bolje rezultate u odnosu na jednostavne modele, potrebno je proceniti njegovu složenost i vreme potrebno za dobijanje rezultata. Takođe, potrebno je napomenuti da su za ovo istraživanje metode za poređenje razvijenog algoritma najjednostavnije. Izbor najjednostavnijih metoda za poređenje temelji se na tome što je poželjno, barem u prvoj iteraciji, dati algoritmu sve moguće uslove da prikaže pozitivan rezultat. Ukoliko to ne bude slučaj, potrebno je vratiti se na početak i razviti novi algoritam primenom drugih metoda.

Priprema i odabir skupa podataka vođeni su sličnim razmišljanjem. U prvoj iteraciji potrebno dati algoritmu sve moguće uslove za postizanje pozitivnog rezultata. Iz baze podataka Agencije za zaštitu životne sredine za period od 2018. do 2021. godine odabrani su podaci koji imaju najduži neprekinuti niz merenih podataka. Ukupno je odabrano 9 stanica, koje se nalaze u Republici Srbiji, a koje sadrže između 4715 konstantno merenih sati podataka (196 merenih dana) i 8,735 konstantno merenih sati (364 konstantno merenih dana). Tri merne stanice se nalaze u Beogradu (Mostar, Novi Beograd, Stari grad), dok Niš (osnovna škola „Sveti Sava“) i Novi Sad (Rumenačka ulica) imaju po jednu mernu stanicu. Pored tri najveća grada u Republici Srbiji, ostale analizirane stanice nalaze se u Obrenovcu, Smederevu, Valjevu i Kosjeriću.

Odabir dužine unetih, nasumičnih preskočenih opservacija u set podataka takođe je pratio isto prethodno razmišljanje. Kao što je prikazano na slici 10, preskočene opservacije u određenim situacijama traju i po nekoliko meseci, što predstavlja veliki problem za istraživače. Sa druge strane, prilikom testiranja algoritma, unete preskočene opservacije u setove podataka imale su dužinu od 24, 48 i 72 sata. Glavna ideja bila je da se dopusti algoritmu da prikaže određene dnevne varijacije u vrednostima za imputaciju, ali da se ograniči dužina preskočenih opservacija kako bi se testirao algoritam sa najjednostavnijim metodama imputacije podataka. Zbog toga se smatra da su tri odabrana intervala preskočenih opservacija kompromis između kvantitativnog i kvalitativnog testiranja algoritma sa najjednostavnijim metodama.

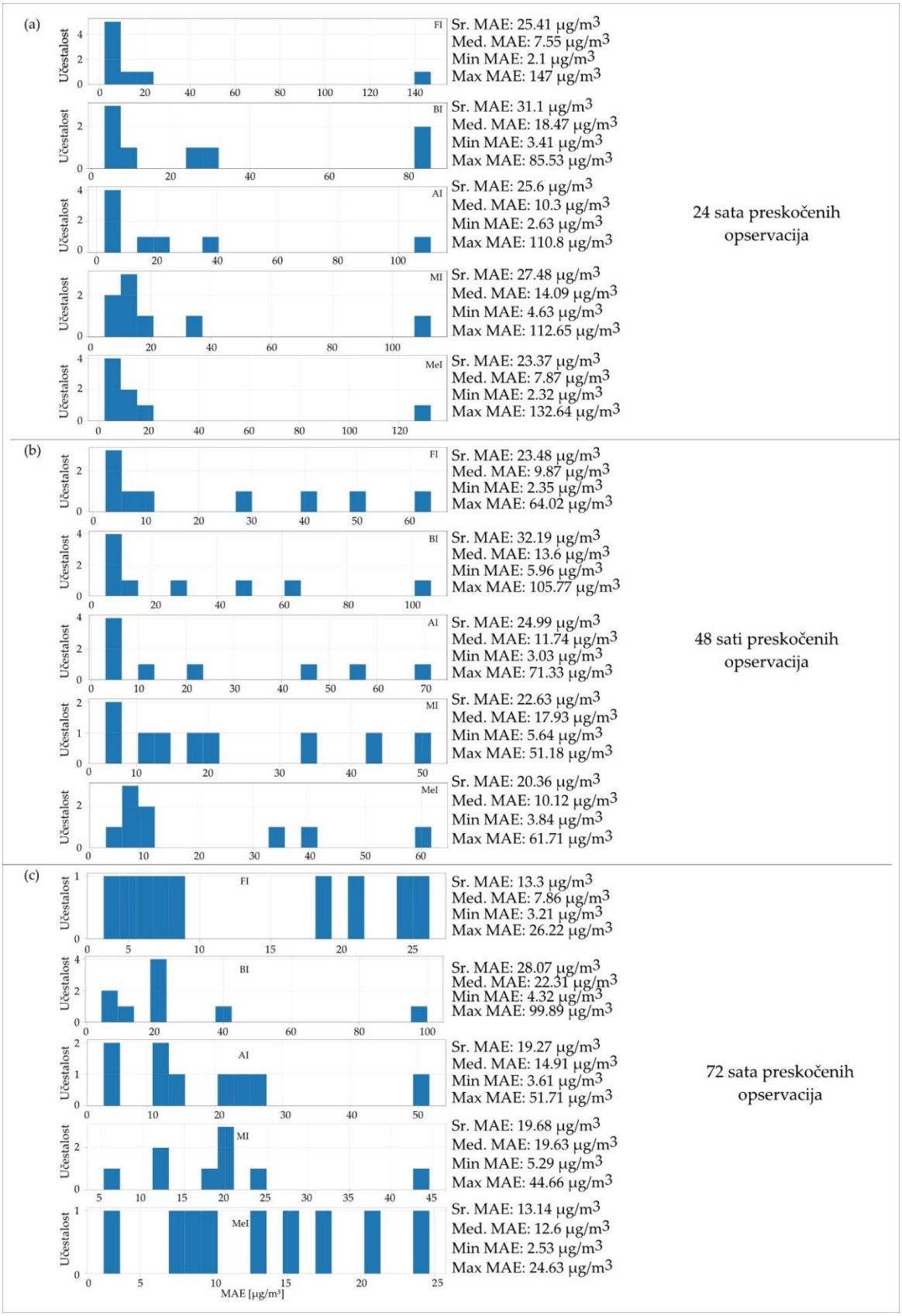
3.1.2.2. Rezultati imputacije podataka

Rezultati imputacije podataka prikazani su na slici 12 za sva tri perioda preskočenih opservacija (24, 48 i 72 sata) i za sve korišćene metode. Sve korišćene metode imputacije podataka prikazale su jedan podatak koji je van okvira drugih podataka. Za imputaciju podataka razvijenim algoritmom smerom unapred (FI) ta vrednost iznosi 140 μg/m³, dok se za druge primenjene modele imputacije ta vrednost nalazi između 80 i 120 μg/m³. Prilikom računanja osrednjenje MAE vrednosti, ovaj podatak unosi asimetriju u raspodelu MAE parametra, čime podiže ukupnu srednju vrednost MAE vrednosti. Prilikom poređenja medijalnih vrednosti, dva modela prednjače: metoda prognoziranja unapred razvijenim algoritmom sa 7,55 μg/m³ i imputacija medijalnom vrednošću sa 7,87 μg/m³. Od svih prikazanih metoda, za raspon preskočenih opservacija od 24 sata, najlošije rezultate dala je metoda imputacije sa razvijenim algoritmom u nazad (18,47 μg/m³).

Analiza skupa podataka koji je prikazao najlošije rezultate za sve korišćene modele pokazala je da set podataka sa najlošijim rezultatima pripada gradu Valjevu. Set podataka pokriva period od 20. juna 2018. godine do 31. marta 2019. godine, odnosno sadrži 6803 konstantno merena sata (283 konstantno merena dana). Prilikom kreiranja skupa podataka za imputaciju, odnosno unošenja nasumičnih 24 sata preskočenih opservacija, prikazano je da se prozor od 24 sata nalazi u danima kada su se vrednosti $PM_{2.5}$ parametra kretale u opsegu od 80 do 250 μg/m³, sa izraženim varijacijama. Zbog velike varijacije u podacima, lokacije preskočenih opservacija (decembar 2018. godine) i temperatura koje su se kretale u vrednostima od 0 °C, lako je razumeti zašto su svi modeli dali loše rezultate. Zbog specifičnosti ovog primera, dodata su dva dodatna modela: prvi je *iterative imputer*, a drugi predstavlja primenu slučajne šume sa atributima (meteorološki atributi, dodatne vrednosti koncentracije zagađujućih materija u vazduhu i dr.). Oba modela su takođe dala loše rezultate, sa vrednostima od 79,7 μg/m³ i 90 μg/m³. Zbog toga, ovaj primer treba uzeti sa rezervom, jer su čak i modeli koji se smatraju vrlo savremenim i adekvatnim za imputaciju preskočenih opservacija dali loše rezultate.

Za primer sa 48 sati preskočenih opservacija, ukoliko se razmatraju srednje vrednosti MAE parametra, imputacija srednjom vrednošću ili medijanom predstavljaju najbolje modele. Sa druge strane, ukoliko se posmatra medijalna vrednost MAE parametra, odabir se svodi na razvijeni algoritam sa imputacijom unapred ili imputaciju medijalnom vrednošću. U ovom delu je potrebno uvesti složenost računanja sa razvijenim algoritmom. Za ceo set test podataka, koji sadrži ukupno 9 stanica raspoređenih po Republici Srbiji, za predloženi algoritam bilo je potrebno u proseku oko 13,21 minuta za izvršenje modelovanja. Imputacija podataka srednjom vrednošću ili medijalnom vrednošću generiše se gotovo trenutno. Ako se kompleksnost računanja uzme kao parametar za razmatranje, vrednosti odstupanja stvarnih i izračunatih vrednosti za predloženi algoritam nisu adekvatne za poređenje, jer predloženi algoritam prikazuje slične vrednosti kao jednostavne metode, ali uz daleko veće računarsko vreme.

Primer za 72 sata preskočenih opservacija slična je kao prethodna dva primera. Imputacija podataka sa smerom unazad uvek prikazuje najlošije rezultate, dok smer unapred i imputacija medijalnom vrednošću prikazuju najbolje rezultate. Ukoliko se razmatraju samo imputacija srednjom vrednošću i medijalnom vrednošću, u svakom analiziranom slučaju medijalna vrednost daje bolje rezultate. Ukoliko je neophodno koristiti neku od najjednostavnijih metoda za imputaciju podataka, imputacija medijalnom vrednošću predstavlja bolji izbor od imputacije srednjom vrednošću, uz jednako utrošeno računarsko vreme.



Slika 12. Rezultati imputacije podataka sa predloženim algoritmom i poređenje sa jednostavnim metodama imputacije podataka

3.2. Primena satelitskih i geofizičkih podataka za klasifikaciju prostornog položaja ofiolita istočne Vardarske zone

U narednom primeru biće prikazana primena modela vođenih podacima na podatke koji su predstavljeni u obliku prostorne serije. Primer primene modela vođenih podacima na prostorne serije biće prikazan kroz primer prostorne klasifikacije ofiolita- delova okeanske kore sa afinitetom srednje- okeanskog grebena ili supra- subdukcione zone, koji su tektonski smešteni na kontinentalnu marginu prilikom zatvaranja okeana. Ofioliti istočne Vardarske zone predstavljaju kompleksnu jedinicu koja se sastoji od metamorfnih stena (gabro- doleriti, doleriti, dajkovi dolerita i dr.), dok se ređe mogu pronaći serpentinisani harzburgiti. Preko slojeva ofiolita uglavnom se nalaze titonski krečnjaci i/ili sedimenti krede (Dimitrijević, 1997). Detaljna sinteza i geološka evolucija istočne Vardarske zone prikazana je u Boev et al. (2018).

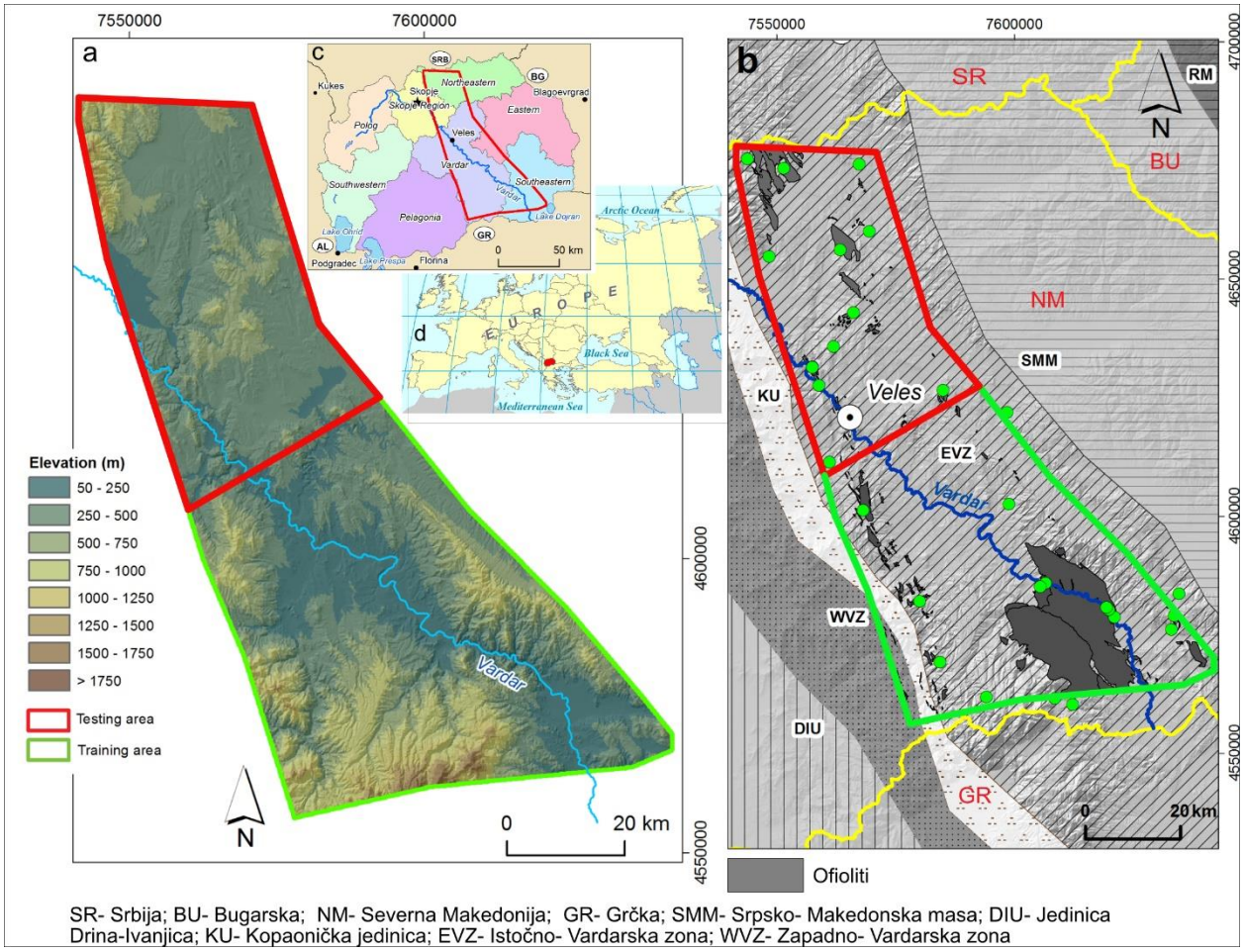
Kombinovanje satelitskih i geofizičkih predstavlja novi pravac u prostornoj klasifikaciji različitih litologija. Takođe, prilikom pretraživanja literature, mogu se naći brojni primeri primene satelitskih podataka za prostornu klasifikaciju geoloških struktura ili litologija, naročito u slučajevima gde je teren slabo prekriven vegetacijom (Aliyu et al., 2021; Lorenz, 2004; Harris et al., 2005, 2008, 2009, 2014; Schetselaar & Ryan, 2008; Leverington, 2010; Leverington & Moon, 2012; Behnia et al., 2012; Harris & Grunsky, 2015; Albert & Ammar, 2021). Sa druge strane, kada je teren prekriven vegetacijom ili sedimentima koji nisu predmet istraživanja, javljaju se problemi prilikom primene metoda mašinskog učenja (Leverington & Moon, 2012; Harris & Grunsky, 2015; Kuhn et al., 2018; Ge et al., 2022). Problemi koji nastaju pri takvoj vrsti klasifikacije, koja se oslanja isključivo na satelitske podatke, koji predstavljaju površinske podatke, ogledaju se u smanjenom kapacitetu modela za klasifikaciju. Naime, ulazni podaci u slučaju terena prekrivenog vegetacijom ili sedimentima ne pružaju informacije o strukturama koje se nalaze ispod njih. Sa druge strane, geofizički podaci, poput gravimetrijskih i magnetometrijskih podataka, sadrže informacije o strukturama ispod površine, što znači da bi kombinacija geofizičkih podataka i satelitskih podataka trebalo da omogući kvalitetan ulaz za klasifikacione modele, bez obzira na pokrivenost strukture ili litologije koja se klasifikuje.

Cilj ovog primera predstavljen je sa dva dela: prvi deo se odnosi na primenu klasifikacionih modela mašinskog učenja za prostornu klasifikaciju ofiolita istočne Vardarske zone, sa

naglaskom na proveru u kojoj meri kombinacija geofizičkih podataka i satelitskih podataka doprinosi uspešnoj klasifikaciji litologije u prisustvu vegetacije. Drugi cilj je da se ispita uticaj balansiranja klasa ciljne promenljive na rezultate prostorne klasifikacije ofiolita, odnosno, u kojoj meri odnos klasa ciljne promenljive utiče na ukupni izlaz modela.

3.2.1. Korišćeni podaci i postavka istraživanja prostorne klasifikacije ofiolita

Istočna Vardarska zona predstavlja istočni deo složene tektonske zone koja se proteže kroz centralni deo Balkanskog poluostrva- Vardarske zone (Karamata, 2006; Schmid et al., 2008, Robertson et al., 2009). Prostire kroz Rumuniju, Srbiju, čitavu Severnu Makedoniju, Grčku i Tursku (Slika 13a). Za potrebe ove disertacije, fokus je na delu istočne Vardarske zone koji se nalazi u Severnoj Makedoniji, a koji se graniči sa Kopaoničkom jedinicom i Srpsko-Makedonskom masom (Slika 13b). Podela istočne Vardarske zone na trening i test set prikazana je na slici 13b, kao i položaj kartiranih ofiolita unutar ove zone u Severnoj Makedoniji, koji u ovom primeru predstavlja ciljnu promenljivu.



Slika 13. (a) Digitalni elevacioni model istočne Vardarske zone; (b) Pojednostavljena geotektonska karta istražnog područja (modifikovano prema Robertson et al., 2009); (c) Položaj istočne Vardarske zone u Severnoj Makedoniji; (d) Položaj Severne Makedonije u sklopu Evropskog kontinenta

Kao što je prethodno spomenuto, ofioliti su ciljna promenljiva, dok su satelitski i geofizički podaci atributi. Tabela 4 prikazuje spisak svih atributa korišćenih u ovom primeru. Satelitski snimci su preuzeti sa Landsat 7 ETM+ sa rezolucijom od 30 metara. Snimci su prošli kroz standardne metode obrade, uključujući geometrijsku i atmosfersku korekciju, filtriranje, povećanje kontrasta i druge tehnike. Takođe, korišćeni su odnosi različitih kanala, koji su pokazali dobre rezultate za kartiranje litologije (Bolt & Bruggenwert 1976; Sposito 1989; Farrand 1997; Longhi et al. 2001; Akhavi et al. 2001; Neville et al. 2003; Al-Rawashdeh et al. 2006), uključujući BR1 (odnos kanala 3/1), BR2 (odnos kanala 5/4) i BR3 (odnos kanala 5/7).

Tabela 4. Prikaz primenjenih atributa za istraživanje podeljeni prema grupi i tipu podataka

Parametar	Simbol	Grupa	Tip podatka	Upotreba	Ostalo
Landsat 7 ETM+ kanal 1	C1	Satelitski	Celobrojni	Atribut	0,45-0,52 μm
Landsat 7 ETM+ kanal 2	C2	Satelitski	Celobrojni	Atribut	0,52-0,60 μm
Landsat 7 ETM+ kanal 3	C3	Satelitski	Celobrojni	Atribut	0,63-0,69 μm
Landsat 7 ETM+ kanal 4	C4	Satelitski	Celobrojni	Atribut	0,77-0,90 μm
Landsat 7 ETM+ kanal 5	C5	Satelitski	Celobrojni	Atribut	1,55-1,75 μm
Landsat 7 ETM+ kanal 7	C7	Satelitski	Celobrojni	Atribut	2,09-2,35 μm
Odnos kanala 1	BR1	Satelitski	Sa ostatkom	Atribut	C3/C1
Odnos kanala 2	BR2	Satelitski	Sa ostatkom	Atribut	C5/C4
Odnos kanala 3	BR3	Satelitski	Sa ostatkom	Atribut	C5/C7
Digitalni elevacioni model	DEM	Geofizički	Celobrojni	Atribut	/
Karta Bugeovih anomalija	BAM	Geofizički	Sa ostatkom	Atribut	/
Totalni intenzitet Zemljinog magnetnog polja nakon redukcije na pol	RTP	Geofizički	Sa ostatkom	Atribut	/
Karta udaljenosti od raseda	DF	Geološki	Sa ostatkom	Atribut	/
Karta položaja ofiolita	Ofioliti	Geološki	Kategorički	Ciljna promenljiva	/

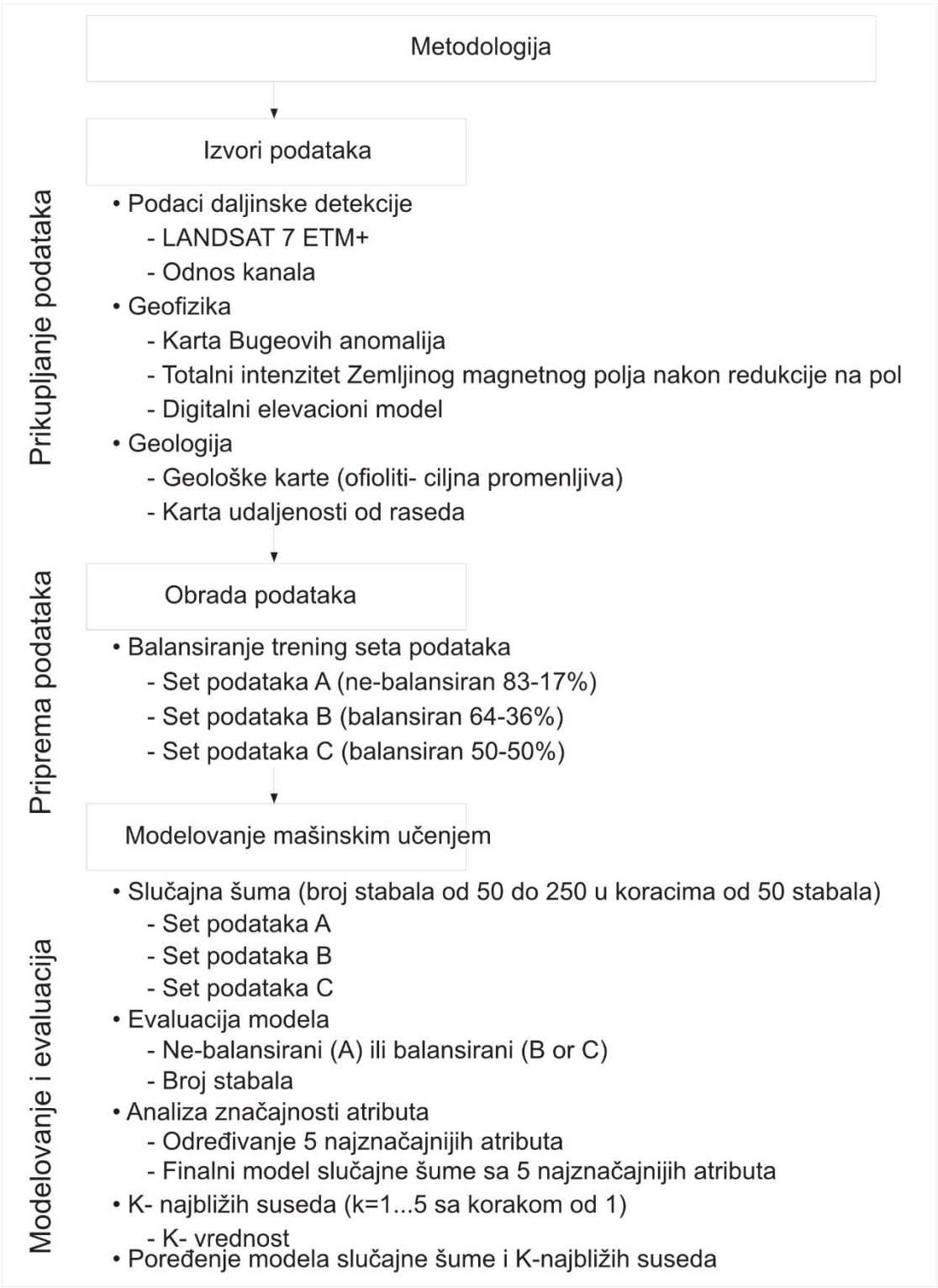
Geofizički podaci predstavljeni su digitalnim elevacionim modelom, kartom Bugeovih anomalija i kartom totalnog intenziteta Zemljinog magnetnog polja nakon redukcije na pol (Tabela 4). Digitalni elevacioni model preuzet je od ASTER misije sa rezolucijom od 30 m. Karta Bugeovih anomalija (Bilibajkić et al., 1979) potiče iz perioda između 1952. i 1984. godine izrađena od strane Geofizičkog instituta. Na podatke je primenjena Kasinisova jednačina (1930), kao i Bulard A korekcija, Bugeova korekcija i korekcija za slobodan vazduh prvog reda. Karta anomalija totalnog intenziteta geomagnetnog polja dobijena je odgovarajućim preračunavanjem karte anomalija vertikalne komponente geomagnetnog polja (Cvetkov et al., 2016). Nad tako dobijenim podacima primenjena je redukcija na pol (Cvetkov et al., 2016), uzimajući u obzir Kenigsebrgov odnos od 0,001 za lokalitet istočne Vardarske zone (Petrović, 2015).

Digitalizacijom osnovnih geoloških karata (Pendžerkovski et al., 1963; Rakićević et al., 1965, 1969, 1973; Ivanovski, Rakicević, 1966; Hristov et al., 1965, 1973; Karajovanović & Hristov, 1976; Dumurdzanov et al., 1981; Dimitrijević, 1978; Karajovanović & Hadži-Mitrova, 1982) dobijena je ciljana promenljiva, kao i poslednji atribut koji predstavlja udaljenost svake

lokacije od raseda na istražnom području. Ciljna promenljiva je predstavljena kao binarna promenljiva, koja označava klasu ofiolita (1) ili klasu svih drugih stena (0). U klasu ofiolita ubrajane su sve stene iz ove grupacije (doleriti, gabro- doleriti, gabrovi, bazaltne pilov lave, serpentinisani harzburgiti), bez obzira na njihovu međusobnu različitost za prvu iteraciju istraživanja, dok je plan za buduća istraživanja da se unutar klase ofiolita izvrši detaljna podela u zavisnosti od sastava.

Radni tok istraživanja prikazan je na slici 14, gde su prethodno prikazani izvori podataka. Nakon prikupljanja svih neophodnih podataka, naredni korak predstavlja pripremu setova podataka. U prvoj iteraciji, set podataka A je kreiran, koji predstavlja nebalansirani set podataka, odnosno onakav kakav je u izvornim podacima. Odnos klasa u ciljnoj promenljivoj je 83- 17% u korist klase "ne ofiolita". Setovi podataka B i C su napravljeni balansiranjem ciljne promenljive, odnosno nasumičnim poduzorkovanjem klase "ne ofiolita". U tom slučaju, napravljena su dva nova seta podataka: set podataka B, sa odnosom klasa 64- 36% u korist klase "ne ofiolita", i set podataka C, sa potpuno balansiranom ciljnom promenljivom.

Modelovanje je izvršeno korišćenjem dva modela: modela slučajnih šuma i K- najbližih suseda. Model slučajnih šuma primenjen je na sva tri seta podataka, sa pretraživanjem broja stabala u opsegu od 50 do 250 u koracima od 50 stabala. Prva evaluacija obuhvatila je poređenje između balansiranih i nebalansiranih podataka, kao i odabir najefikasnijeg broja stabala. Analiza značajnosti atributa sprovedena je radi identifikovanja 5 najinformativnijih atributa, što je omogućilo odabir finalnog modela sa smanjenim brojem atributa u odnosu na prvobitni model. Model K- najbližih suseda rađen je sa vrednostima K od 1 do 5, sa korakom od 1. Za model K- najbližih suseda primenjen je umanjen broj atributa, odnosno 5 najinformativnijih atributa iz modela slučajnih šuma. Na kraju, izvršeno je poređenje između modela slučajnih šuma i K- najbližih suseda.



Slika 14. Radni tok istraživanja prostorne klasifikacije ofiolita istočne Vardarske zone

Tabela 5 prikazuje raspodelu podataka u trening i test setovima za sve tri grupe podataka korišćene u ovom istraživanju. Set podataka A sadrži najveći broj podataka u trening setu, sa ukupno 328 hiljada podataka i odnosom između klasa 83- 17%. Set podataka B sadrži upola

manje podataka u trening setu (155 hiljada), ali je odnos klasa bolje balansiran u odnosu na inicijalni set, sa 64- 36%. Set podataka C ima najmanji broj podataka u trening setu, sa ukupno 110 hiljada podataka i potpuno izbalansiranim klasama ofiolita i ne ofiolita. Test set podataka je isti za sva tri seta, sa oko 200 hiljada podataka i velikim disbalansom između klasa ofiolita i ne ofiolita (93- 7%).

Tabela 5. Raspodela klasa ciljne promenljive u tri seta podataka za prostornu klasifikaciju ofiolita istočne Vardarske zone

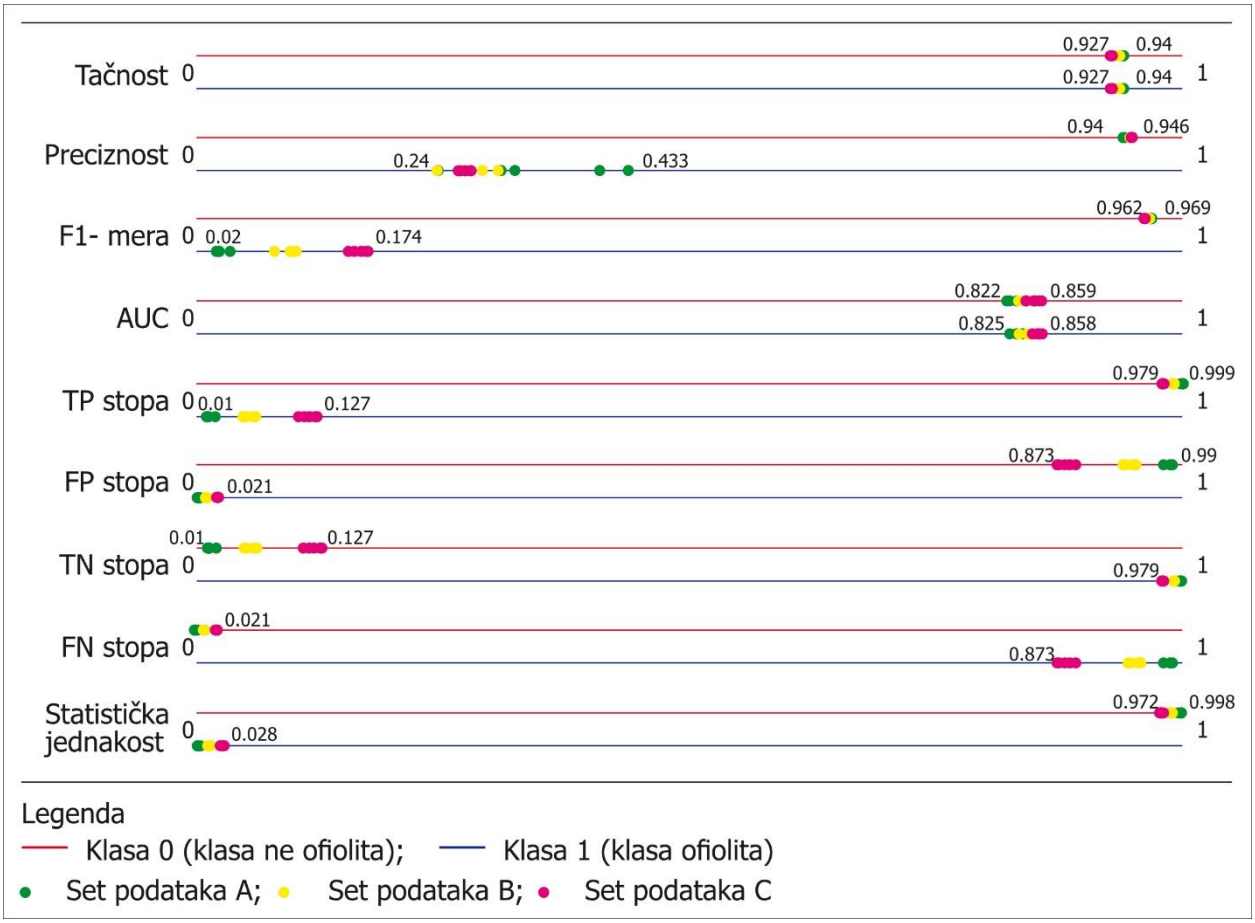
Set podataka A					
	0 [/]	**1 [/]**	**Suma [/]**	**0 [%]**	**1 [%]**
Trening	273 106	55 454	328 560	83%	17%
Testing	189 116	12 138	201 254	93%	7%
Suma	462 222	67 592	529 814	87,20%	12,80%
Set podataka B					
	0 [/]	**1 [/]**	**Suma [/]**	**0 [%]**	**1 [%]**
Trening	100 000	55 454	155 454	64,30%	35,70%
Testing	189 116	12 138	201 254	93%	7%
Suma	289 116	67 592	356 708	81%	19%
Set podataka C					
	0 [/]	**1 [/]**	**Suma [/]**	**0 [%]**	**1 [%]**
Trening	55 454	55 454	110 908	50%	50%
Testing	189 116	12 138	201 254	93%	7%
Suma	244 570	67 592	312 162	78%	22%

0- klasa ne ofiolita; **1** - klasa ofiolita;
Sum - Ukupan broj podataka po setu podataka

3.2.2. Rezultati modelovanja vođenih podacima za klasifikaciju prostornog položaja ofiolita

Prva iteracija rezultata dobijenih modelovanjem vođenim podacima prikazana je na slici 15, gde se može uočiti da za parametar tačnosti postoji vrlo mala razlika između sva tri seta podataka, koja ne prelazi 1,3%. Parametar preciznosti pokazuje značajne razlike između tri seta, pri čemu set podataka A ima najveće vrednosti. Sa druge strane, F1-mera za klasu ofiolita, koja je ključni parametar, pokazuje razliku od 15,4% u korist seta podataka C. Svi prikazani modeli imaju zadovoljavajuće vrednosti AUC parametra, koji se kreću u opsegu od 82,2% do 85,9%. Stopa tačno pozitivnih instanci pokazuje najveće vrednosti kod seta podataka C sa 12,7%, dok najniže vrednosti ima set podataka A. Na osnovu prikazanih kvantitativnih mera

kvaliteta modela, može se zaključiti da je set podataka C najbolji od tri pripremljena seta podataka za postupak sprovedenog modelovanja.



Slika 15. Odabrane mere kvantitativne ocene kvaliteta modela klasifikacije

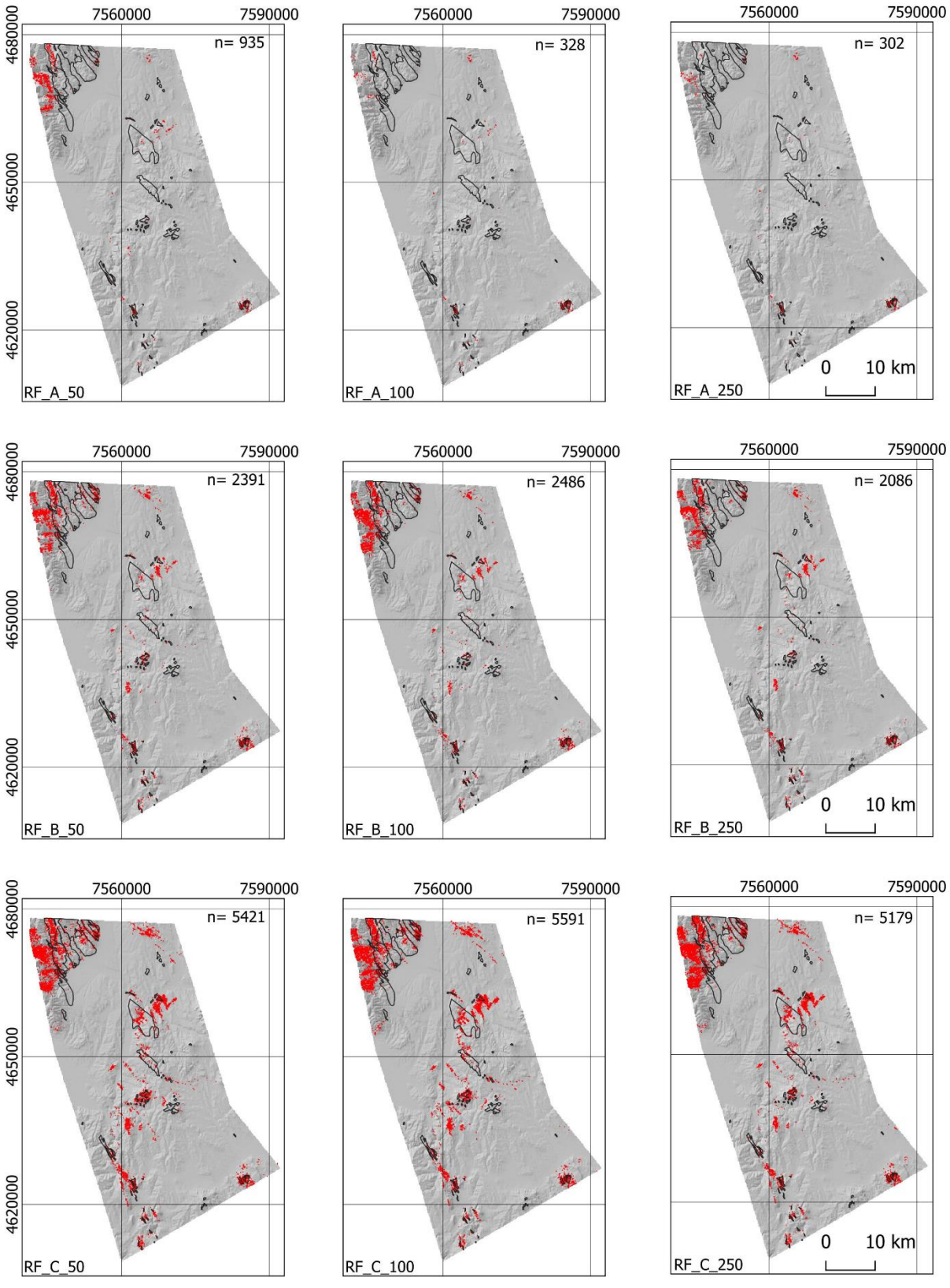
Drugi vid evaluacije modela prostorne klasifikacije ofiolita može se izvršiti prikazivanjem instanci koje je model klasifikovao kao klasu ofiolita na test prostoru. Slika 16 prikazuje rezultate modela slučajnih šuma koji su koristili 50, 100 i 250 stabala, primenjenih na sva tri seta podataka (A, B i C). Kao što je prethodno prikazano, set podataka A pokazuje najlošiji balans između klase ofiolita i neofiolita u trening setu podataka (83- 17%), dok su druga dva seta podataka bolje izbalansirana. Zbog toga, model slučajnih šuma sastavljen od seta podataka A prikazuje najmanji broj instanci koje je klasifikovao kao klasu ofiolita (935 za 50 stabala i oko 300 za 100 i 250 stabala). Sa druge strane, set podataka sa potpuno uravnoteženim klasama ofiolita i neofiolita prikazuje najveći broj instanci klasifikovanih kao ofioliti, oko 5 hiljada. Slično daje i parametar statističke jednakosti prikazanim na slici 15, gde set podataka C pokazuje najveću statističku jednakost za klasu ofiolita. Takođe, potrebno je naglasiti da je Tabela 5 pokazala da set podataka A sadrži otprilike tri puta više podataka u trening setu u

poređenju sa setom podataka C, ali je set podataka C pokazao najbolje rezultate među sva tri seta. Drugim rečima, i sa tri puta manjim brojem podataka, balansiranjem klasa ciljne promenljive moguće je postići bolje klasifikacije.

Prethodno je prikazano da je model slučajnih šuma, korišćenjem seta podataka C, najbolji, a unutar svih napravljenih modela (5), postoji vrlo mala razlika u parametrima evaluacije modela. Slična situacija je i sa kartom raspodele klasifikovanih instanci ofiolita na test području, gde su razlike između modela minimalne, pa je odabran model sa 100 stabala.

Takođe, potrebno je interpretirati relativno malu vrednost F1- mere za klasu ofiolita na test području. Vrednosti F1- mere iznose 17,4%, što nije velika vrednost i ukazuje na potrebu za daljim unapređenjem modela. Ipak, važno je napomenuti da je model u dobroj meri prikazao prostornu logiku rasporeda ofiolita na test području. Naime, ofioliti se nalaze na tri grupisane lokacije. Prva lokacija je na severozapadnom delu test područja, druga, manja lokacija se nalazi na jugoistočnom delu, dok treća lokacija čini pojas koji se proteže od jugozapada do severoistoka. Klasifikovane instance modela tačno prikazuju te lokacije, iako nisu sve instance ofiolita klasifikovane ispravno (Slika 16).

Analiza značajnosti atributa pokazala je da su pet najinformativnijih atributa: karta Bugeovih anomalija, karta udaljenosti od raseda, digitalni elevacioni model, karta totalnog intenziteta Zemljinog magnetnog polja i odnos kanala 3 (BR3). Sa druge strane, među atributima sa najmanjom informativnošću za model nalaze se kanal 4 Landsat satelitskih snimaka, odnos kanala BR1 i BR2, kao i ostali kanali Landsat satelitskih snimaka.

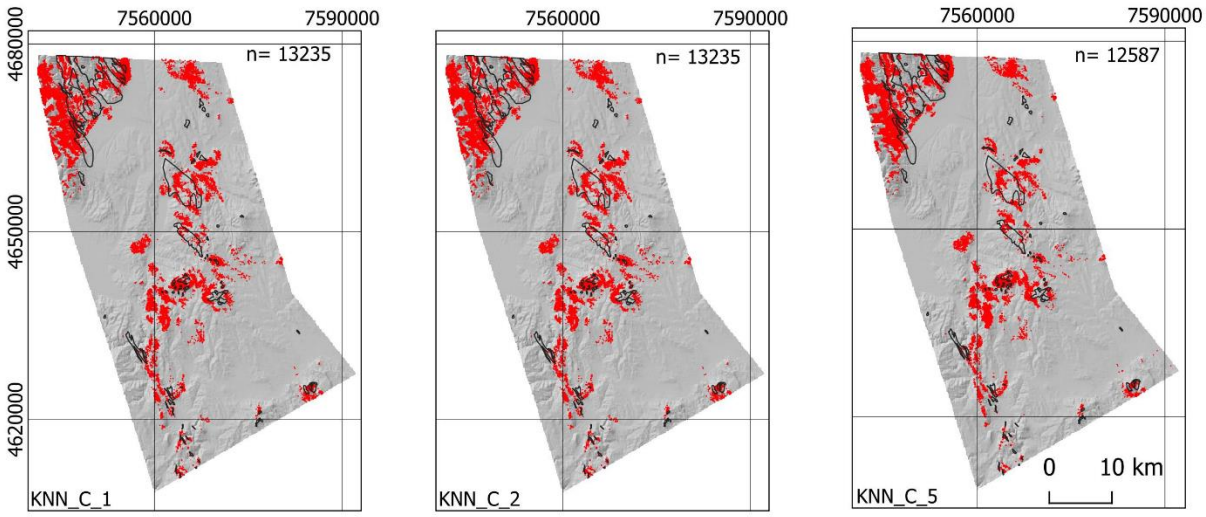


Slika 16. Primer klasifikacije za model slučajne šume; (a) Nebalansirani- A set podataka; (b) Balansirani- B set podataka i (c) Balansirani- C set podataka; Crno- kartirani ofioliti; Crveno- klasifikovane instance ofiolita od strane modela

Prilikom konstrukcije modela sa samo 5 najinformativnijih atributa, koji su prethodno prikazani, dobijene su relativno uporedive vrednosti kvantifikacije modela klasifikacije. Na primer, parametar tačnosti kod modela sa svim atributima iznosi 92,7% za klasu ofiolita, dok je kod modela sa samo 5 najinformativnijih atributa 92,2%. Sa druge strane, veće razlike prikazuju parametri preciznosti (27,4% i 18,2%) i F1- mere (17,4% i 11,8%) za klasu ofiolita, dok je parametar AUC pokazao slične vrednosti (85,2% i 83,1%). Takođe, prilikom ove analize i smanjenja broja atributa, važno je uporediti i utrošeno računarsko vreme. U ovom slučaju, kada se posmatra razlika u utrošenom računarskom vremenu (210 i 93 sekunde) nije značajna, ali razlika od 55% manje utrošenog računarskog vremena može, za veća istražna područja sa većim brojem atributa i možda više klasa ciljne promenljive, doneti značajne uštede na računarskom vremenu, uz minimalno žrtvovanje kvaliteta modela.

Nastavak istraživanja predstavlja primenu modela K- najbližih suseda sa pet najinformativnijih atributa, koji su prethodno korišćeni u modelu slučajnih šuma. Poređenje modela slučajnih šuma i K- najbližih suseda prikazuje najveću razliku u parametru statističke jednakosti. Naime, kod modela slučajnih šuma ovaj parametar iznosi 2,8%, dok kod modela K- najbližih suseda vrednost iznosi oko 6,6%, što je bliže stvarnoj raspodeli klasa u test setu podataka (93- 7%). Drugi značajan uvid iz parametra statističke jednakosti je da model K- najbližih suseda klasifikuje oko dva i po puta više instanci kao ofiolite. Parametar F1- mere za model K- najbližih suseda varira od 17,6% do 19,3%, što je uporedivo sa vrednostima parametra slučajnih šuma. Parametar AUC, sa druge strane, pokazuje niže vrednosti, u opsegu od 56,3% do 65,1%. Preciznost modela K- najbližih suseda se kreće u opsegu od 16,8% do 18,9%, što je manji opseg u poređenju sa modelom slučajnih šuma. Međutim, ove vrednosti nisu uporedive sa vrednostima sa Slike 15, jer su na slici 15 prikazani i setovi podataka A i B, koji nisu testirani kod modela K- najbližih suseda.

Slika 17 prikazuje klasifikacije modela K- najbližih suseda, gde se može uočiti da, kao što se i očekivalo, ovaj model klasifikuje veći broj instanci kao klasu ofiolita, sa oko 13 hiljada instanci, dok je model slučajnih šuma klasifikovao oko 5 hiljada instanci kao ofiolite. Takođe, kao i kod modela slučajnih šuma, iako model K- najbližih suseda nije klasifikovao veliku većinu instanci ofiolita kao ofiolite, on je ipak prikazao dobru prostornu raspodelu klasifikovanih instanci, koje se uklapaju u opšti trend prostorne raspodele ofiolita na test području.

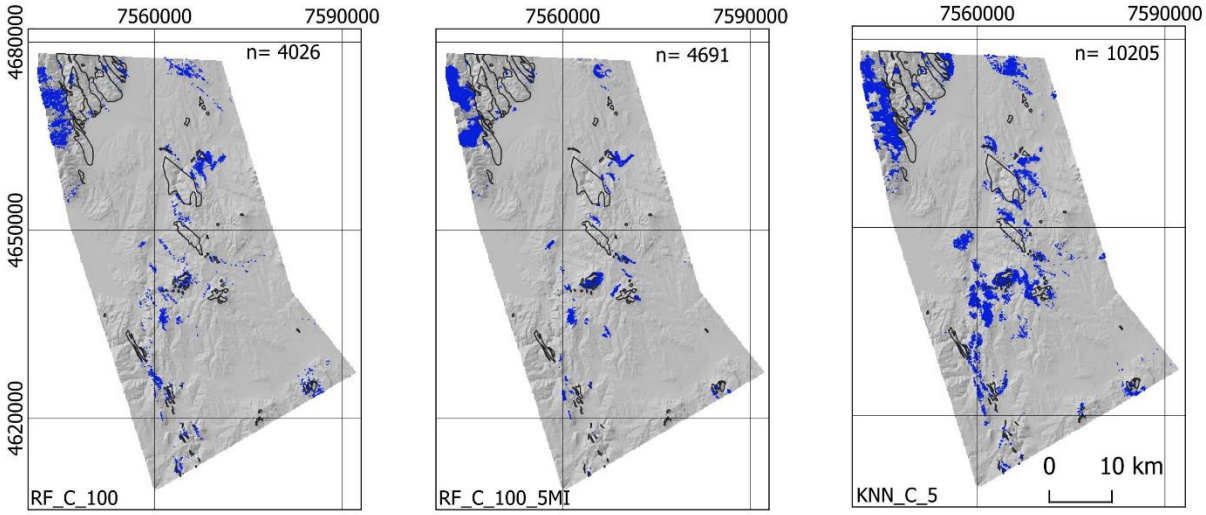


Slika 17. Prostorne klasifikacije modela K- najbližih suseda za vrednosti K parametra 1- 3
Crno- kartirani ofioliti; Crveno- klasifikovane instance ofiolita od strane modela

Prilikom poređenja modela slučajne šume i K- najbližih suseda, prethodno je prikazano da je model slučajnih šuma postigao bolje rezultate, što je naročito vidljivo u parametru AUC. Drugi način da se prikažu razlike između ova dva modela predstavlja analiza karti lažno pozitivnih instanci (Slika 18). Može se uočiti da model K- najbližih suseda klasifikuje duplo više instanci kao lažno pozitivne u poređenju sa modelom slučajnih šuma, i to kako za pun set atributa, tako i za set sa 5 najinformativnijih atributa. Takođe, razlika u FP stopi za klasu ofiolita između modela slučajnih šuma (2,1%) i modela K- najbližih suseda (5,8%) je očigledna. Drugim rečima, iako model K- najbližih suseda prikazuje relativno dobru raspodelu klase ofiolita (6,6%) koja je bliska stvarnoj raspodeli klase ofiolita u test setu podataka (7%), veliki broj tih instanci je klasifikovan kao lažno pozitivan.

Model K- najbližih suseda, kao što je prethodno prikazano, pokazuje najveću klasifikacionu moć uz mali broj atributa. Zbog toga je odlučeno da se u ovom primeru koristi model K- najbližih suseda sa samo pet najinformativnijih atributa, jer se očekivalo da će ovaj pristup dati najbolje rezultate. Sa druge strane, model slučajnih šuma pokazuje izuzetnu sposobnost modelovanja kompleksnih i nelinearnih odnosa između atributa i ciljne promenljive, te je zbog toga model slučajnih šuma sa 100 stabala i 5 najinformativnijih atributa izabran kao najbolji i finalni model. Model sa 5 najinformativnijih atributa i 100 stabala predstavlja kompromis između utrošenog računarskog vremena, lakog proširivanja i adaptiranja za dalja istraživanja i

dobrih kvantitativnih vrednosti mera modela, uz odličnu prostornu raspodelu klasifikovanih ofiolita.



Slika 18. Lažno pozitivne instance; (a) Model slučajnih šuma sa potpunim setom atributa; (b) Model slučajnih šuma sa pet najinformativnijih atributa i (c) Model K- najbližih suseda sa K-vrednosti od 5; Crno- kartirani ofioliti; Plavo- lažno pozitivne instance od modela

Prikazane lažno pozitivne instance na test setu podataka, pored poređenja između modela i odabira najboljeg finalnog modela, pružaju još jednu vrlo bitnu informaciju. Kao što je prethodno objašnjeno, satelitski podaci su površinski podaci, tj. ne sadrže informacije o podpovršinskoj građi. Sa druge strane, primenjeni geofizički podaci (karta Bugeovih anomalija i karta totalnog intenziteta Zemljinog magnetnog polja nakon redukcije na pol) nose informacije o strukturama u podpovršini.

Prilikom konstrukcije ciljne promenljive korišćene su osnovne geološke karte koje prikazuju kartirane ofiolite sa površine terena. Na taj način, modelu je data ciljna promenljiva koja je površinski vezana, dok su atributi sadrže informacije o strukturama u podpovršini (geofizički podaci) i atributi koji su takođe površinski vezani (satelitski podaci). Drugim rečima, model ne prepoznaje da li su atributi i ciljna promenljiva povezani sa površinom ili dubinom. Poznato je da je ciljna promenljiva površinski vezana jer je kartiranje izvršeno sa površine terena.

Zbog ovoga, karta lažno pozitivnih instanci se može drugačije interpretirati. Lažno pozitivne instance mogu ukazivati na lokacije u test setu podataka koje nemaju ofiolite na površini terena,

ali ih ima u podpovršini. U tom slučaju, prema lažno pozitivnim instancama, mogu se rasčlaniti dve grupe:

- Stvarno lažno pozitivne instance- Lokacije u kojima ofioliti nisu prisutni ni na površini, ni na određenoj dubini i
- Prividno lažno pozitivne instance- Lokacije u kojima ofioliti nisu prisutni na površini, ali se nalaze na određenoj dubini i nisu kartirani.

Raščlanjivanje stvarno i prividno lažno pozitivnih instanci zahteva terenski rad i verovatno određena istražna bušenja, što može biti ekonomski neisplativo. Sa druge strane, karta lažno pozitivnih instanci može biti koristan ulaz za dalja istraživanja i potvrdu od strane drugih disciplina koje se bave tektonikom, strukturom i genezom istočne Vardarske zone ukoliko se model u daljem istraživačkom radu unapredi.

Da bi se ispitalo da li je moguće unaprediti prostornu klasifikaciju ofiolita, model koji je prethodno davao najbolje rezultate (slučajne šume) zamenjen je modelom ekstremnog povećanja gradijenta. Takođe, prostor pretraživanja hiperparametara je proširen: broj stabala je povećan na interval od 50 do 1000 stabala sa korakom od 10. Funkcija pretraživanja hiperparametara promenjena je i sada je predstavljena nasumičnim pretraživanjem hiperparametara, sa ukupno 100 modela. Drugi ključni hiperparametar modela ekstremnog povećanja gradijenta, stopa učenja, definisan je u opsegu od 0,005 do 0,3, sa korakom od 0,005, takođe uz nasumično pretraživanje hiperparametara. Najznačajnije unapređenje u ovoj iteraciji istraživanja predstavlja dodatak novog atributa, nazvanog "*karta udaljenosti od reke*". Podaci o prostornoj raspodeli reka na trening i test području pretvoreni su u kartu (digitalizovanjem), koja je zatim dodata kao novi atribut modelu.

Ukupno su izvršena dva testa kako bi se kvantifikovalo koliko svaki od dva prethodno prikazana noviteta doprinosi unapređenju modela. Prvi test se sastoji u zameni modela slučajnih šuma modelom ekstremnog povećanja gradijenta, bez dodavanja atributa karte udaljenosti od reka. Drugi test uključuje dodavanje pomenutog atributa na već postojeći model ekstremnog povećanja gradijenta, nakon čega se vrši poređenje između ocena kvaliteta modela. Tabela 6 prikazuje rezultate modela ekstremnog povećanja gradijenta bez dodatka atributa karte udaljenosti od reka, kao i modela ekstremnog povećanja gradijenta sa dodatim atributom karte udaljenosti od reka. Poređenje modela bez atributa karte udaljenosti od reka sa modelom slučajnih šuma koji je koristio ceo set atributa pokazuje relativno slične vrednosti preciznosti

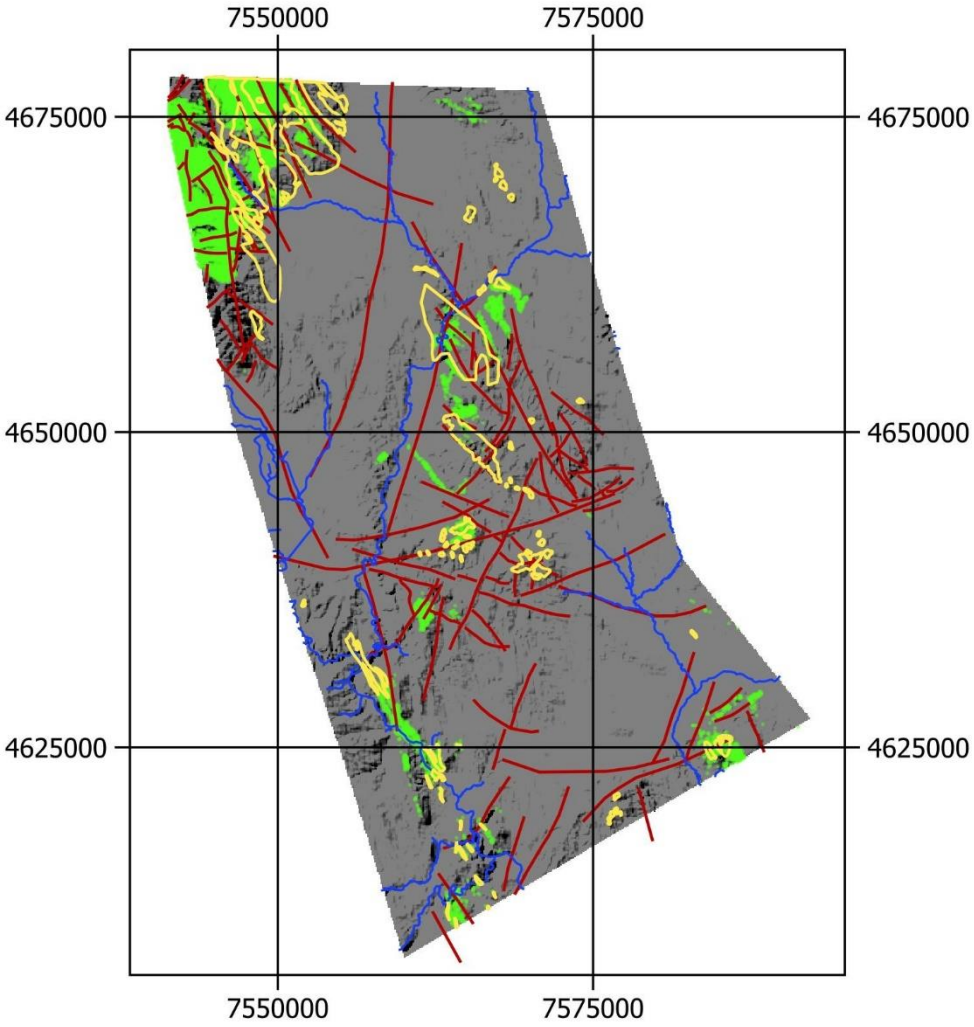
za klasu ofiolita (27,4% i 26%), ali i povećanje stope tačno pozitivnih instanci za 12,3% (sa 12,7% na 25%) i poboljšanje F1- mere za klasu ofiolita sa 17,4% na 25%. Samo zamena modela slučajnih šuma modelom ekstremnog povećanja gradijenta, uz unapređenje pretraživanja hiperparametara, već je dala značajno poboljšanje u rezultatima. Sa druge strane, kada se atribut karte udaljenosti od reka doda modelu, poboljšanja su još izraženija, uključujući 5,6% povećanje preciznosti, 16,3% povećanje stope tačno pozitivnih instanci i 13,6% poboljšanje F1- mere za klasu ofiolita.

Tabela 6. Kvantifikacija modela iz druge iteracije istraživanja

Model/Klasa	Preciznost		Stopa TP		F1- mera	
	0	1	0	1	0	1
Bez atributa udaljenosti od reka	0,95	0,26	0,95	0,25	0,95	0,25
Sa atributom udaljenosti od reka	0,95	0,33	0,96	0,29	0,96	0,31

Slika 19 prikazuje prostorni položaj istinitih (kartiranih) ofiolita, prostorni položaj klasifikovanih ofiolita, kao i položaj reka i trase raseda na test području. Kao i u prethodnim primerima, model je odgovarajuće prikazao prostornu raspodelu ofiolita na test području istočne Vardarske zone. Model je tačno prikazao da se ofioliti nalaze u tri grupe: severozapadni deo, centralni pojas i mali deo jugoistočnog dela test područja. Pored poboljšanja kvaliteta modela, ostaje objašnjenje za razlog tog poboljšanja. U prvoj iteraciji odgovor je jasan- promena modela i proširenje prostora pretraživanja hiperparametara dovelo je do početnog povećanja kvaliteta modela.

Drugi korak u poboljšanju kvaliteta modela rezultat je uvođenja atributa karte udaljenosti od reka. Geološko objašnjenje za ovo poboljšanje može se naći u činjenici da se ofioliti Ždraljice i Kuršumlije uglavnom nalaze u blizini reka. Prisustvo izdanaka ofiolita u blizini reka može se povezati sa erozijom koju su reke vršile tokom vremena, uklanjajući površinske slojeve, dok su ofioliti ostali skriveni, slično se pretpostavlja i na primeru iz Severne Makedonije. Sa stanovišta mašinskog učenja, jedno od mogućih objašnjenja je da je kompleksna i najverovatnije nelinearna kombinacija svih atributa, uključujući atribut karte udaljenosti od reka, omogućila modelu da postigne značajno bolju sposobnost klasifikacije u odnosu na prethodni model, čime je efikasnije razlikovao ofiolite od drugih litologija na test setu podataka.



Slika 19. Klasifikacija ofiolita druge iteracije istraživanja; Žuto- kartirani ofioliti; Zeleno- klasifikovani ofioliti od strane modela: Crveno- trasa raseda; Plavo- trasa reka

3.3. Primena metoda mašinskog učenja na podatke signala vrlo niske frekvencije (VLF) koji se prostiru sub- jonosferski

Niska jonosfera prostire se u rasponu visina od 50 do 90 km iznad Zemljine površi (Fedrizzi et al., 2002) i predstavlja zonu koja varira u jonizaciji pod uticajem Sunca, kao i u skladu sa efektima solarnih flerova (eng. *Solar flares*) (Mitra, 1978; Ohya et al., 2006; Kumar et al., 2014; Ahmedov et al., 2020). Pored efekata solarnih flerova, na jonosferu takođe utiču izbacivanja koronalne mase. Moderni komunikacioni sistemi, kao što su sateliti, navigacioni sistemi i drugi radio signali, prostiru se u rasponu visina koja odgovaraju jonosferi. Zbog toga, uticaj ekstraterestričkih pojava na jonosferu može imati značajan efekat na komunikacione sisteme.

Značaj primene modela vođenih podacima za podatke niske jonosfere ogleda se u dva ključna primera. Prvi primer predstavlja primenu modela mašinskog učenja za klasifikaciju i detekciju anomalija na signalu vrlo niske frekvencije (VLF). Ovaj primer je značajan jer primena metoda mašinskog učenja omogućava klasifikaciju nepoželjnog VLF signala. Drugi primer se odnosi na upotrebu mašinskog učenja za određivanje parametara niske jonosfere, kao što su visina reflektovanja i oštrina granice reflektovanja VLF signala prilikom njihove sub-jonosferske propagacije. Za modelovanje ovih parametara standardno se koristi vrlo kompleksna metoda, te primena mašinskog učenja predstavlja potencijalnu alternativu koja zaslužuje dodatnu istraživanja.

U oba prethodno navedena primera, jedan od glavnih uzročnika poremećaja niske jonosfere su solarni flerovi. Solarni flerovi predstavljaju kratkotrajne, vrlo intenzivne izbačaje elektromagnetnog zračenja sa Sunčeve površine. Zračenje stigne do Zemljine okoline u vrlo kratkom vremenskom periodu (reda veličine do nekoliko minuta), dok u slučaju izbacivanja solarne koronalne mase (eng. *Coronal mass ejections*), materijal stiže u dužem vremenskom periodu, obično u trajanju od nekoliko dana.

Solarni flerovi se klasifikuju prema intenzitetu, pri čemu se prema konvenciji intenzitet solarnog flera označava klasom (slovna oznaka), pored koje se nalazi broj koji preciznije označava njegovu jačinu. Tabela 7 prikazuje klase solarnih flerova prema intenzitetu X-zračenja, gde su solarni flerovi X- klase najintenzivniji i generalno najređi, solarni flerovi M-

i C- klase su slabiji i češće se javljaju, dok su flerovi B- klase najslabiji (ako se izuzme A-klasa).

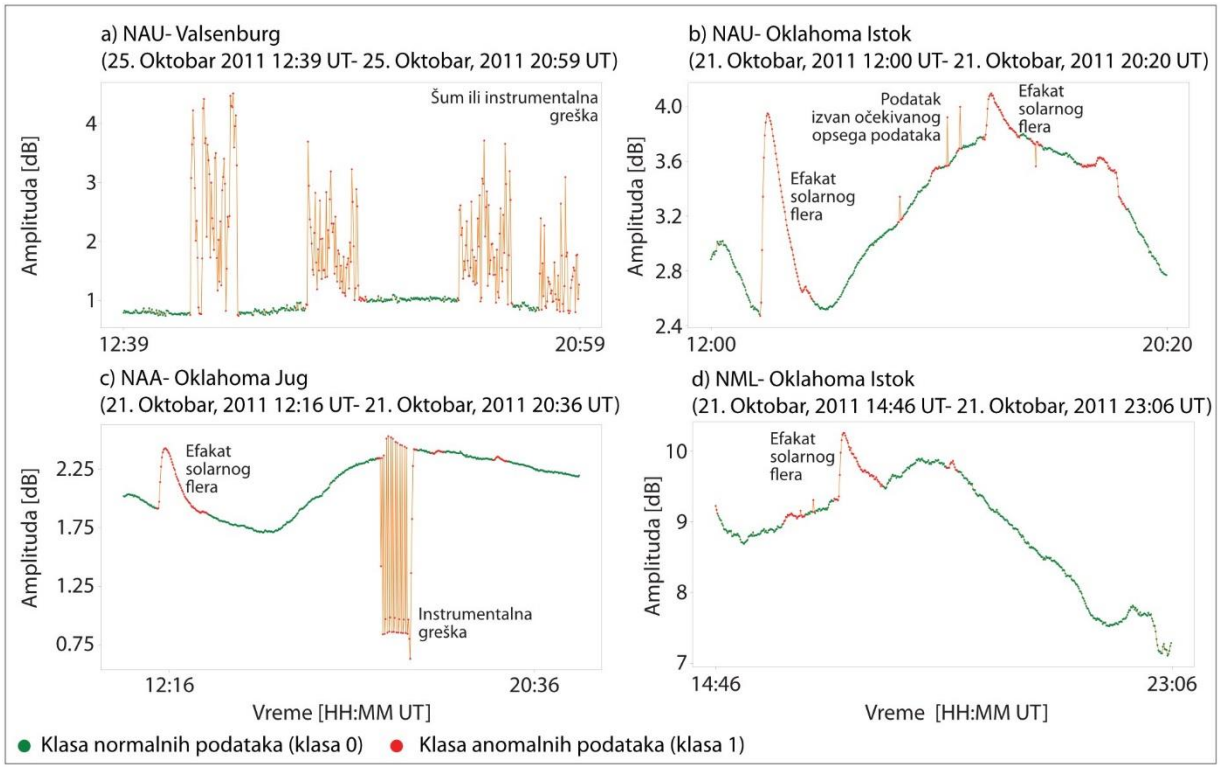
Tabela 7. Klase solarnih flerova prema intenzitetu X- zračenja

Klasa solarnog flera	Intenzitet [Wm^{-2}]
X	$\geq 10^{-4}$
M	$10^{-5} \leq 10^{-4}$
C	$10^{-6} \leq 10^{-5}$
B	$\leq 10^{-6}$
A	$\leq 10^{-7}$

Gornja granica za intenzitet solarnih flerova ne postoji; neki od najjačih zabeleženih solarnih flerova uključuju X40+ koji se dogodio 4. novembra 2003. godine, X28,57 koji se dogodio 2. aprila 2001. godine i X24,57 koji se dogodio 28. oktobra 2003. godine. Verovatnoća za nastanak velikog solarnog flera varira sa solarnim ciklusom (eng. *Solar cycle*), koji je definisan brojem Sunčevih pega i prikazuje cikličnost koja traje oko 11 godina. Tokom solarnog maksimuma, verovatnoća za pojavu jakog solarnog flera je veća, ali treba napomenuti da postoje izuzeci, da su se neki od najjačih solarnih flerova dogodili van perioda maksimuma solarnog ciklusa, kao što su dva pomenuta solarna flera iz 2003. godine, koja su se dogodila na silaznoj grani solarnog ciklusa 23, dok je solarni fler iz 2001. godine javio tokom solarnog maksimuma.

3.3.1. Primena metoda mašinskog učenja za detekciju anomalije amplitude jonosferskog VLF signala

Kao što je prethodno prikazano, prvi primer se odnosi na primenu mašinskog učenja za detekciju anomalnog VLF signala. Pod terminom „anomalni“ VLF signal podrazumeva se svaki signal koji odstupa od normalnog, dnevnog VLF signala. U ovu kategoriju spadaju instrumentalne greške ili šum na signalu (Slika 20a, c), efekti solarnih flerova (Slika 20b, c, d), podaci koji odstupaju izvan očekivanog opsega (eng. *outlier data points*) (Slika 20b), kao i noćni VLF signal. U prvoj iteraciji istraživanja, sve ove grupe se svrstavaju u anomalnu klasu, dok normalni dnevni signal pripada normalnoj klasi VLF podataka.



Slika 20. Primer anomalnih podataka na signalu amplitude VLF signala

Svrha ovog istraživanja je prvenstveno razvijanje algoritma za automatsko „čišćenje“ podataka. Čišćenje VLF signala od neželjenih vrednosti je vremenski intenzivan proces koji istraživač mora obaviti ručno. Razvoj algoritma koji može da razlikuje normalni od anomalnog signala značajno bi uštedeo vreme i trud istraživača. Takođe, ako se razvije višeklasni algoritam, moguće je izdvojiti specifične delove signala, kao što su solarni flerovi ili noćni VLF signali, za dalju analizu. Pored toga, postoji i mogućnost razvoja algoritma koji bi mogao da detektuje različite karakteristike VLF signala u (skoro) realnom vremenu što bi potencijalno bilo od velikog značaja, sa obzirom da se trenutno sve analize i dalje rade na "starim" signalima.

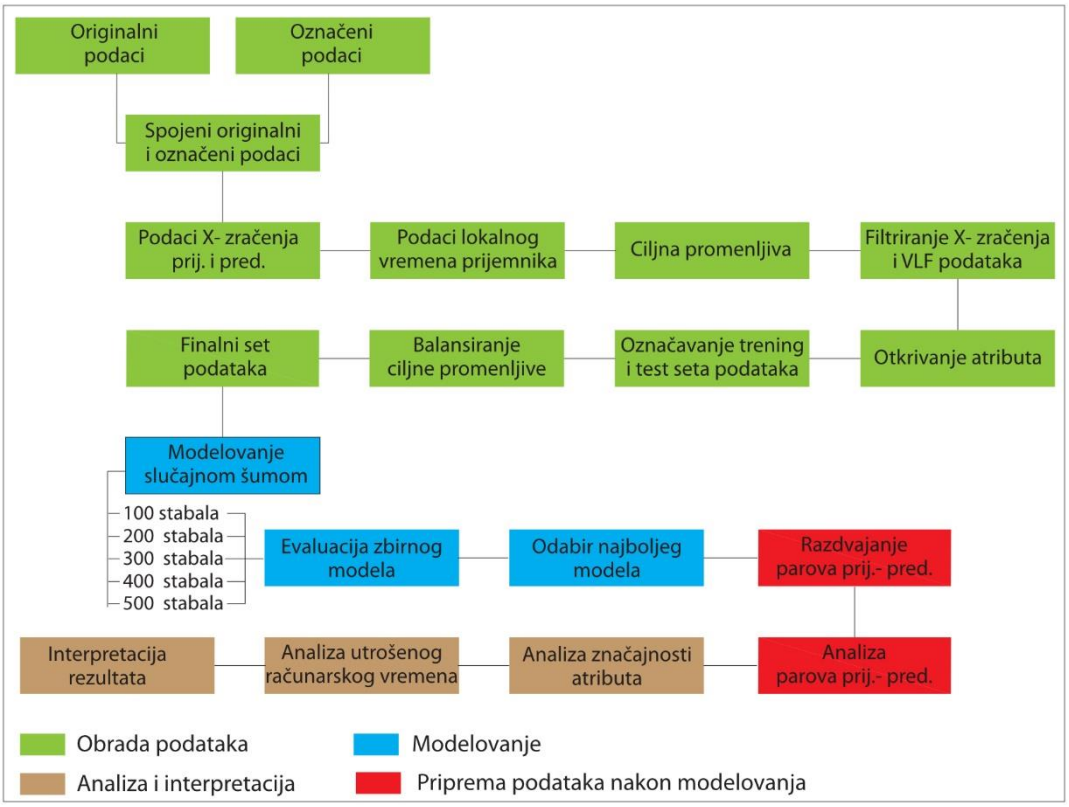
3.3.1.1. Postavka istraživanja, radni tok algoritma i korišćeni podaci

Podaci korišćeni za ovaj primer prikupljeni su u formi vremenskih serija, pri čemu je zavisna varijabla predstavljena amplitudom VLF signala. U ovom istraživanju korišćeno je 5 VLF predajnika: NPM, NLK, NML, NAA, i NAU, te 4 prijemnika: Valsenburg (eng. *Walsenburg*), Oklahoma istok (eng. *Oklahoma East*), Oklahoma jug (eng. *Oklahoma South*) i Šeridan (eng. *Sheridan*). Ukupno je analizirano 19 parova predajnik- prijemnik. Vremenski period analize podataka obuhvata septembar 2011. godine, kada su zabeleženi solarni flerovi klasa C2,5 do X2,1, koji su korišćeni za trening set, dok je test set obuhvatio podatke iz oktobra 2011. godine, sa solarim flerovima klasa C5,5 do M1,5.

Podaci su prethodno ručno označeni i filtrirani za potrebe šireg istraživanja, što je omogućilo da se, bez potrebe za ponovnim obeležavanjem, filtrirani i nefiltrirani podaci spoje u jedinstvenu bazu podataka (Slika 21). Za modelovanje mašinskim učenjem, u ovu kombinovanu bazu podataka dodati su podaci o X- zračenju, lokalnom vremenu prijemnika i definisana ciljana promenljiva. Ciljna promenljiva je predstavljena binarnom klasom: podaci koji su se pojavili u obe grupe (filtrirani i nefiltrirani) predstavljaju normalnu klasu VLF signala (klasa 0), dok podaci koji su bili prisutni u originalnim podacima, a preskočeni u filtriranim, predstavljaju anomalnu klasu (klasa 1).

Filtriranje podataka o X- zračenju i amplitudi VLF signala omogućilo je verifikaciju da nisu ostali nerealni podaci (npr. -9999 ili 999) koji u sebi nose informaciju o instrumentalnim greškama kao i o izostanku registracija. Otkrivanje atributa u ovom slučaju obuhvatilo je primarne atribute, kao što su amplituda VLF signala i X- zračenje, dok su sekundarni atributi uključivali lokalno vreme prijemnika, kao i kodirane informacije o prijemnicima i predajnicima. Tercijarni atributi su obuhvatali izvedene vrednosti kao što su klizne srednje vrednosti, standardna devijacija, medijana, prethodne vrednosti signala, stopa promene, prvi i drugi diferencijal, kao i kategoričke vrednosti koje prikazuju da li je podatak veći od srednje vrednosti ili medijalne vrednosti.

Za klizne vrednosti korišćeni su prozori od 5, 20 i 180 minuta, dok su prethodni podaci obuhvatili intervale od 1 do 5 minuta. Poslednji koraci uključivali su označavanje trening i test podataka, kao i balansiranje ciljne promenljive pomoću metode nasumičnog poduzorkovanja. Modelovanje je vršeno upotrebom modela slučajnih šuma, sa maksimalnim brojem stabala od 500, minimalnim od 100, i korakom od 100 stabala. Nakon modelovanja, izvršena je evaluacija zbirnog modela, dok su zatim analizirani pojedinačni parovi predajnik- prijemnik. Pored toga, urađene su analize utrošenog računarskog vremena i značajnosti atributa.



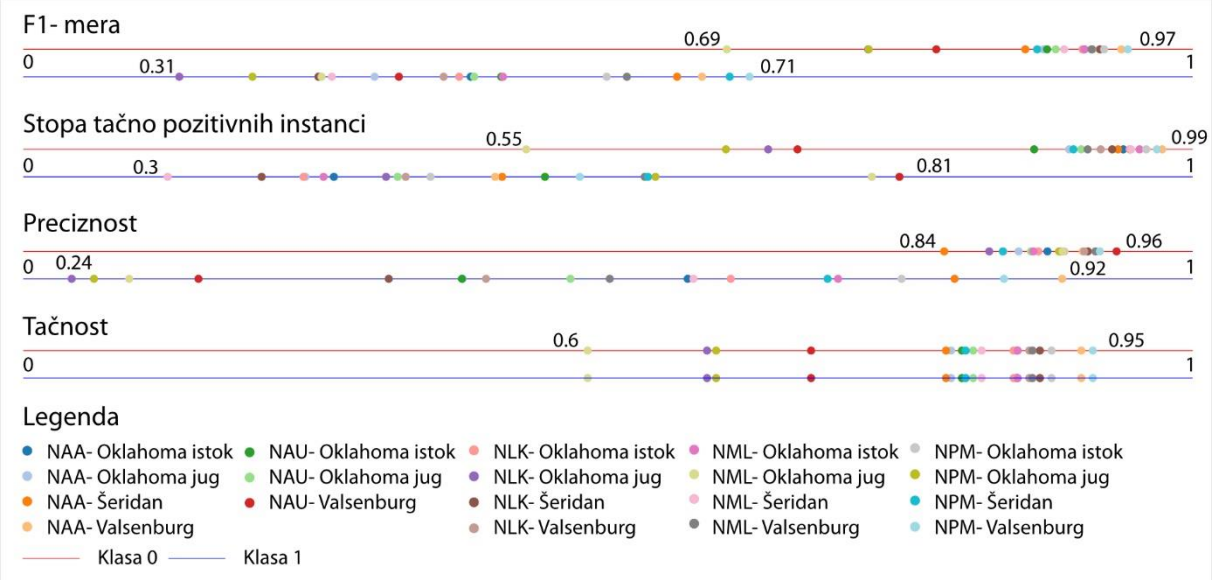
Slika 21. Radni tok modelovanja za detekciju anomalije na signalu amplitude VLF signala

3.3.1.2. Rezultati klasifikacije amplitude jonosferskog VLF signala

Izlaz zbirnog modela testiranog na test setu podataka prikazuje sledeće rezultate: preciznost od 84%, F1- meru od 84,5% i AUC parametar od 84,6%. Zbirni model je pokazao da su modeli generalno dobro trenirani i da razlika između broja stabala nije značajna. Iako zbirni model, koji uključuje podatke svih 19 parova predajnik- prijemnik, daje solidne zbirne rezultate, on ne pruža dovoljno informacija o odstupanjima pojedinačnih parova predajnik- prijemnik, zbog čega je potrebna detaljnija analiza.

Detaljnija analiza je omogućena kroz pojedinačne parove predajnik- prijemnik, što je omogućilo da se identifikuju parovi za koje model daje tačne klasifikacije, kao i oni za koje nije. Na slici 22 prikazane su vrednosti F1- mere za klasu anomalnih vrednosti, koje variraju od 0,31 do 0,71. U svim prikazanim slučajevima (sa Slike 22), jasno je da model bolje klasifikuje klasu normalnih podataka nego klasu anomalnih podataka. Sa obzirom na izraženu razliku u raspodeli klasa u test setu (85% za normalne vrednosti i 15% za anomalne), modelu je teže da tačno klasifikuje anomalni signal.

Takođe, primer sa slike 22 pokazuje da je neophodno analizirati pojedinačne slučajeve kako bi se dobio detaljan uvid u sposobnost modela da klasifikuje anomalni VLF signal, što će biti prikazano u narednim primerima.



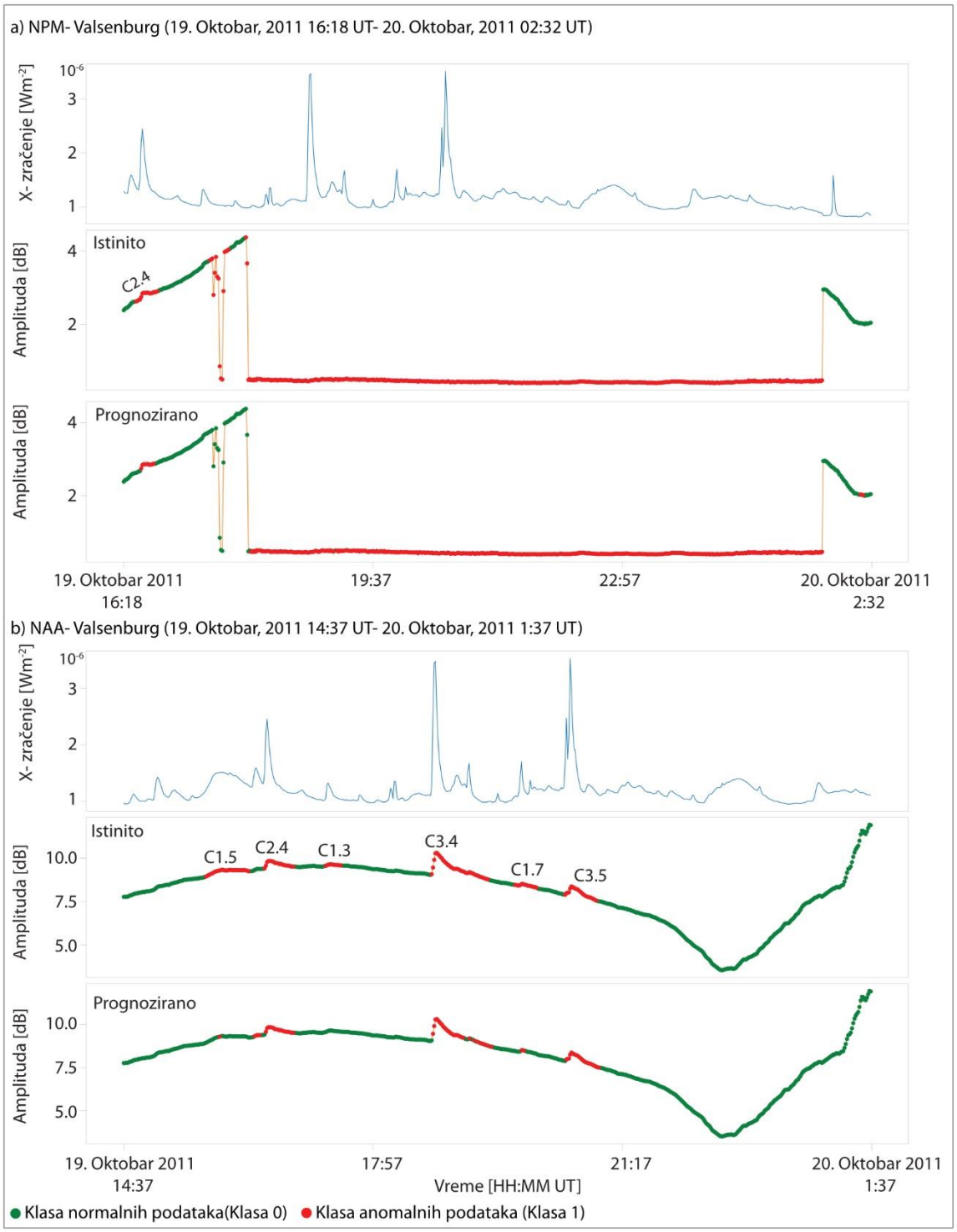
Slika 22. Odabrane mere kvantifikacije modela klasifikacije prema pojedinačnim parovima predajnik- prijemnik

Primer dobre klasifikacije podataka prikazan je na slici 23, gde je klasifikovano nekoliko karakterističnih delova amplitude VLF signala. Slika 23a prikazuje dve instrumentalne greške. Prva greška se manifestuje kroz nasumično merene vrednosti u trajanju od 6 minuta, gde se vrednosti nasumično variraju. Klasifikacija modela za te vrednosti nije bila uspešna, jer ih je model klasifikovao kao normalni VLF signal. Druga greška javlja kao vrednosti signala koje su jednake nuli, što je model adekvatno prepoznao kao anomalni signal.

Na početku signala sa Slike 23a vidi se efekat solarnog flera klase C2,4, koji je model uspešno klasifikovao. Na kraju signala, model je klasifikovao deo normalnog signala kao anomalni signal kratkog trajanja. Takve greške nisu od velikog značaja, jer će u budućim fazama razvoja metode biti moguće filtrirati kratkotrajne lažno pozitivne vrednosti primenom klaster analize (eng. *Cluster analysis*).

Slika 23b prikazuje primer uticaja šest uzastopnih solarnih flerova u opsegu od klasa C1,3 do C3,5. Tri najjača flera (klase C3,5, C3,4 i C2,4) su uspešno klasifikovana, dok su manji solarni flerovi, poput klasa C1,3, C1,5 i C1,7, bili preskočeni ili samo delimično klasifikovani.

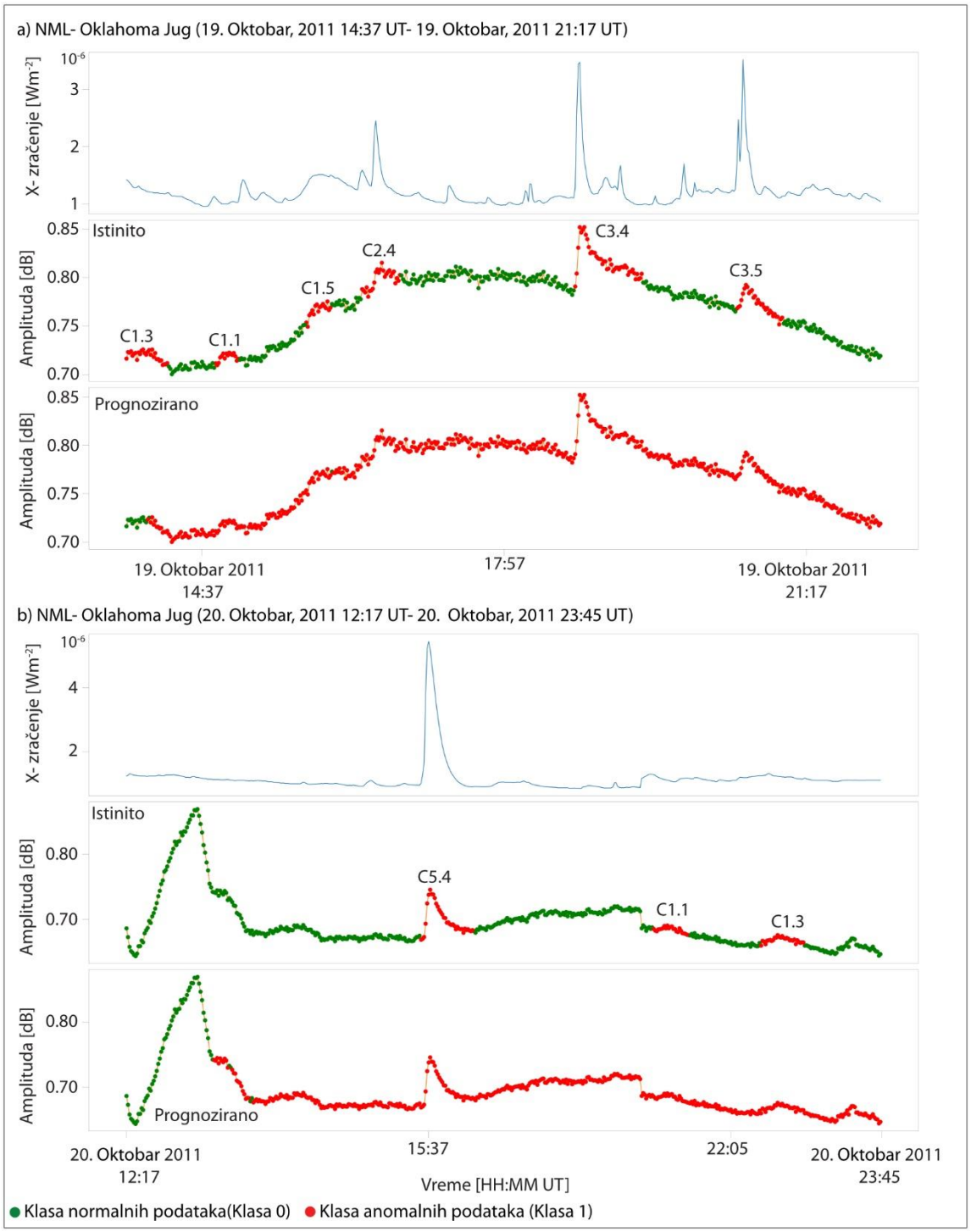
Ovaj rezultat pokazuje da model ima veću uspešnost u klasifikaciji jačih solarnih flerova, što je i očekivano. Jaki solarni flerovi uzrokuju značajne promene u amplitudi VLF signala i X-zračenju jer su visokog X- zračenja, što je i vidljivo na Slici 23b, gde model uspešno klasifikuje tri najjača solarna flera koja uzrokuju velike promene u oba signala.



Slika 23. Primeri zadovoljavajuće klasifikacije modela; (a) Par predajnik- prijemnik NPM-Valsenburg; (b) Par predajnik- prijemnik NAA- Valsenburg

Takođe, bilo je i slučajeva gde je model prikazao vrlo loše klasifikacije amplitude VLF signala, kao što je prikazano na slici 24. U prvom primeru (Slika 24a), moguće je videti uticaj ukupno šest solarnih flerova u opsegu klasa od C1,1 do C3,5. Model je klasifikovao skoro celu trasu signala kao anomalnu, iako je jasno vidljivo na ručno klasifikovanom signalu postojanje šest individualnih solarnih flerova, koji su međusobno razdvojeni normalnim signalom. Važno je napomenuti da je količina smetnji na ovoj trasi velika, što je verovatno predstavljalo problem modelu u tačnoj klasifikaciji.

Situacija na slici 24b je bolja što se tiče šuma, iako šum i dalje postoji, on je manje izražen nego na slici 24a. Na ovom primeru se jasno vide tri solarna flera, od kojih su dva slabijeg intenziteta (C1,1 i C1,3), dok je jedan solarni fler jačeg intenziteta (C5.4). Kao i u prethodnom slučaju, model je klasifikovao skoro celu trasu kao anomalnu. Ukoliko se posmatra X- zračenje za dva solarna flera slabijeg intenziteta, može se primetiti da su odstupanja gotovo neprimetna na trasi X- zračenja, pa loše označavanje tog dela signala nije bilo iznenađujuće. Međutim, označavanje anomalnog signala klase C5,4, koji ima izražena odstupanja u X- zračenju, nije bilo očekivano.

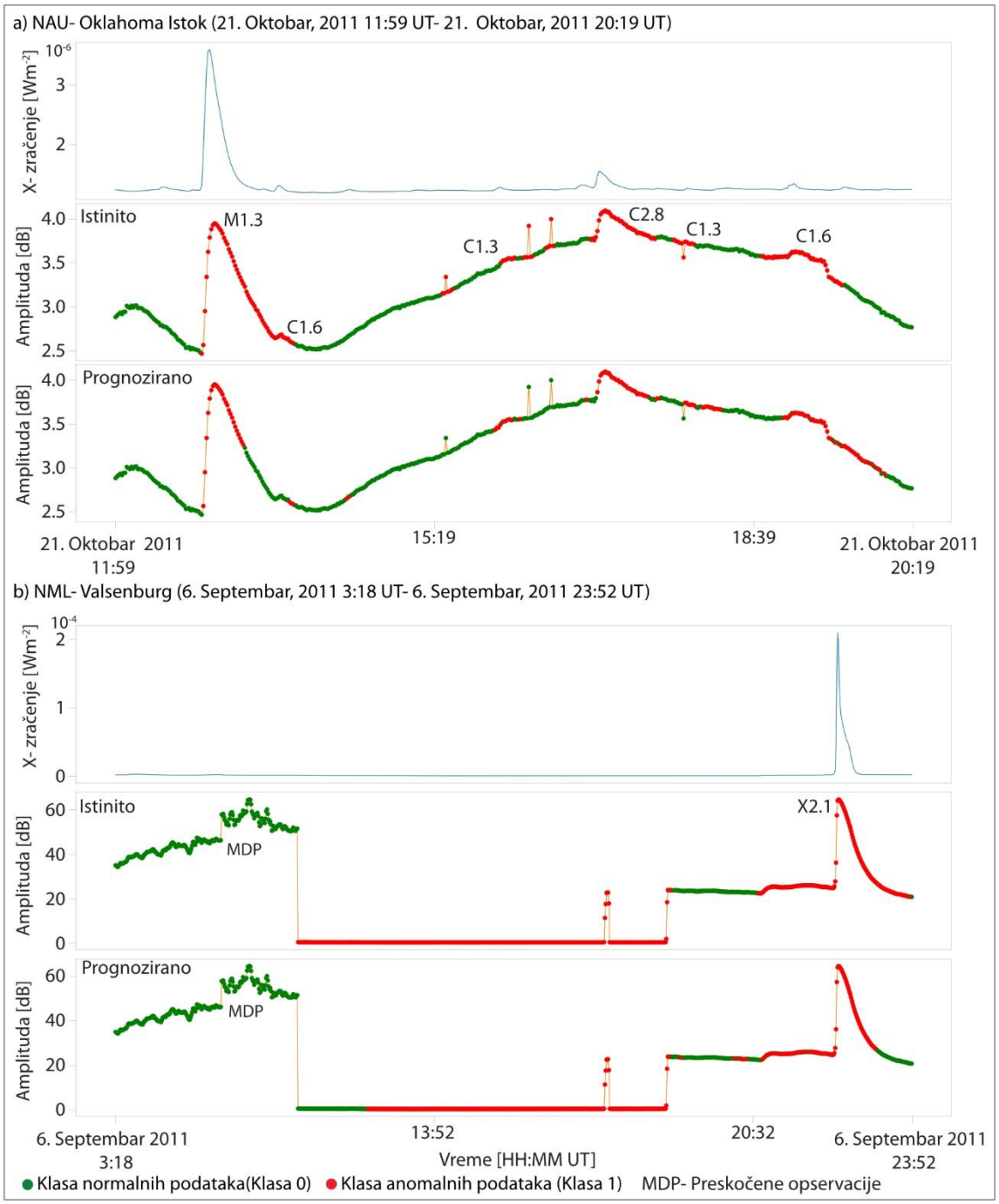


Slika 24. Primeri loše klasifikacije modela; (a) Par predajnik- prijemnik NML- Oklahoma Jug; (b) Par predajnik- prijemnik NML- Oklahoma Jug

Primeri sa Slike 25 prikazuju interesantne slučajeve koji obuhvataju kombinaciju jačih solarnih flerova sa podacima koji izlaze iz okvira ostalih podataka (Slika 25a) ili kombinaciju jakih solarnih flerova sa instrumentalnim greškama (Slika 25b). Solarni fler klase M1,3, kod kojeg se u fazi smirivanja aktivnosti javio solarni fler slabog intenziteta klase C1,6, prikazan je kao delimično dobro klasifikovan na slici 25a. Nedostatak ove klasifikacije je u tome što je silazna

grana solarnog flera klase M1,3 klasifikovana samo do polovine kao anomalni signal, dok je drugi deo klasifikovan kao normalan signal. Drugi najjači solarni fler sa slici 25a (C2,8) je u potpunosti klasifikovan, kao i solarni fler klase C1,6 na kraju signala (iako je dužina anomalnog signala trebala biti kraća od prikazane). Individualni podaci koji izlaze van okvira susednih podataka, kojih ima četiri na slici 25a, su većinom loše klasifikovani kao normalan signal.

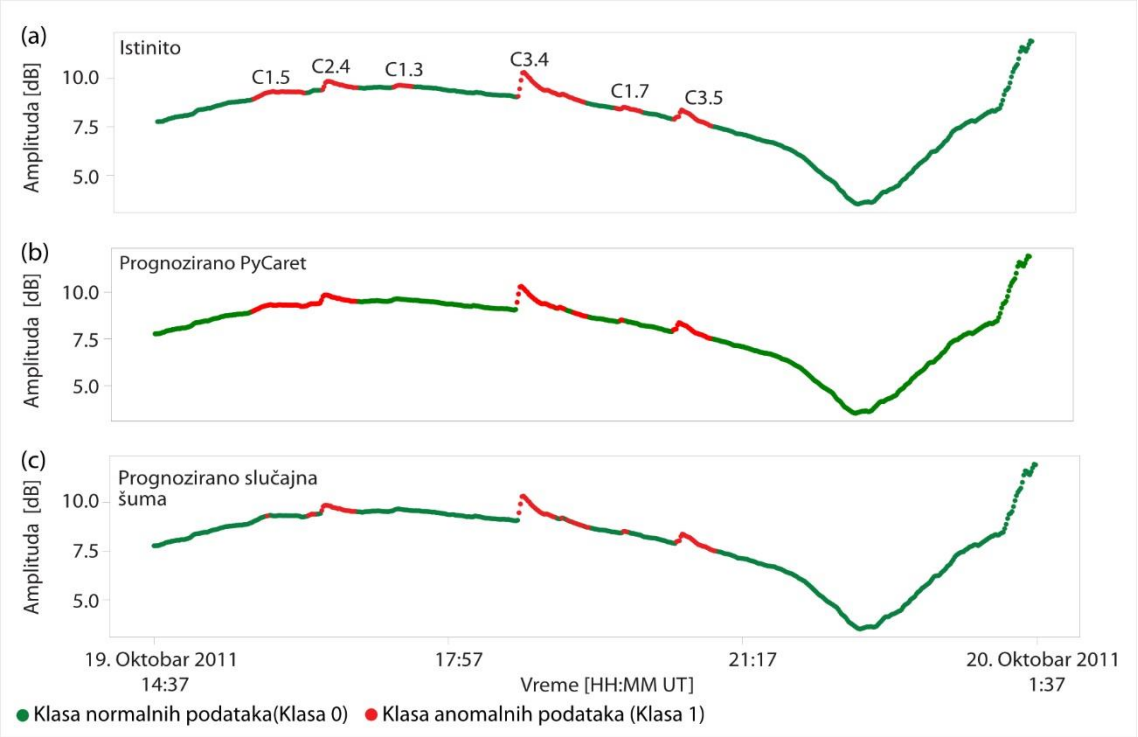
Slika 25b prikazuje kombinaciju vrlo jakog solarnog flera klase X2,1 sa preskočenim opservacijama, odnosno instrumentalnim greškama. Za potrebe ovog primera, trening i test podaci su bili obrnuti kako bi se omogućio primer klasifikacije solarnog flera X- klase. Kao što se može videti, klasifikacija takvog solarnog flera bila je zadovoljavajuća, kao i klasifikacija instrumentalne greške, iako instrumentalna greška nije odmah klasifikovana kao anomalni signal. Par delova klasifikacije prikazuje kratkotrajne klasifikacije normalnog signala kao anomalni signal, ali takvi ostaci se mogu u budućnosti filtrirati klaster analizom.



Slika 25. Primer klasifikacije jačih solarnih flerova zajedno sa podacima koji odstupaju od susednih podataka i instrumentalne greške; (a) Par predajnik- prijemnik NAU- Oklahoma Istok; (b) Par predajnik- prijemnik NML- Valsenburg

Slika 26 prikazuje tri sekcije: istinite klasifikacije, klasifikacije dobijene pomoću PyCaret biblioteke i klasifikacije dobijene modelom slučajnih šuma, kao što je prikazano na primeru sa Slike 23b. Kao što je prethodno pomenuto, na Slici 23b postoji šest jedinstvenih solarnih flerova, a na slici 26 model slučajnih šuma je tačno klasifikovao samo tri, dok su preostali klasifikovani loše. Model odabran pomoću PyCaret biblioteke pokazuje blago poboljšanje u

klasifikaciji. Naime, solarni flerovi klase C1,5 i C2,4 sa početka trase su u modelu PyCaret klasifikovani kao jedna instanca anomalnog signala, dok su u modelu slučajnih šuma klasifikovani samo kao instanca C2,4 sa vrlo kratkom klasifikacijom C1,5. Oba modela prikazuju potpuno preskakanje solarnog flera klase C1,3, ali tačno klasifikuju solarne flerove klasa C3,4 i C3,5. Takođe, oba modela gotovo u potpunosti preskaču solarni fler klase C1,7.



Slika 26. Primer primene biblioteke niskog koda mašinskog učenja na podatke amplitude VLF signala; (a) Istinita klasifikacija; (b) Klasifkovane vrednosti bibliotekom PyCaret i (c) Klasifikovane vrednosti metodom slučajne šume

3.3.2. Primena metoda mašinskog učenja za aproksimaciju talasovodnih parametara oblasti D jonosfere

Prilikom korišćenja tehnologije upotrebe VLF signala transmitovanih talasovodom Zemlja-jonosfera za istraživanje i monitoring niske jonosfere Zemlje, oblast D jonosfere je karakterisan sa dva parametra: oštrinom (β- km^{-1}) i visinom granice reflektovanja signala (*H'*- km). Ovi parametri su takođe poznati kao Vajtovi parametri (eng. *Wait parameters*) ili talasovodni parametri (eng. *Waveguide parameters*) (Wait & Spies, 1964). Koncentracija elektrona može se izračunati za poremećene i neporemećene uslove primenom jednačine uvedene od strane Vajta i Spajsa (eng. *Wait and Spies equation*). Parametri visine i oštrine oblasti D jonosfere tradicionalno se dobijaju korišćenjem LWPC softverskog paketa (eng. *Long Wavelength Propagation Capability*) (Ferguson, 1998), što u uslovima poremećene

jonosferske plazme predstavlja izazov prilikom modelovanja zbog kompleksnih procesa i proračuna koji su sastavni deo pomenutog programa. Razvijene su i različite metode za procenu parametara oblasti D jonosfere bez primene LWPC softverskog paketa, kao što su FlareED i EasyFit (Srećković et al., 2019; Srećković et al., 2021), koje se zasnivaju na aproksimaciji parametara oblasti D jonosfere. Sa druge strane, primena LWPC softvera tokom vrlo poremećenih uslova, kao što je solarni fler klase X17.2 koji se dogodio 28. decembra 2003. godine, može biti dodatno otežana zbog geometrije trase, ograničenja samog modela i drugih restrikcija veznih za putanju po velikom krugu (eng. *Great Circle Path*). Zbog toga se smatra da je razvoj alternativnih metoda koje nisu direktno povezane sa LWPC softverom od velikog značaja za istraživače.

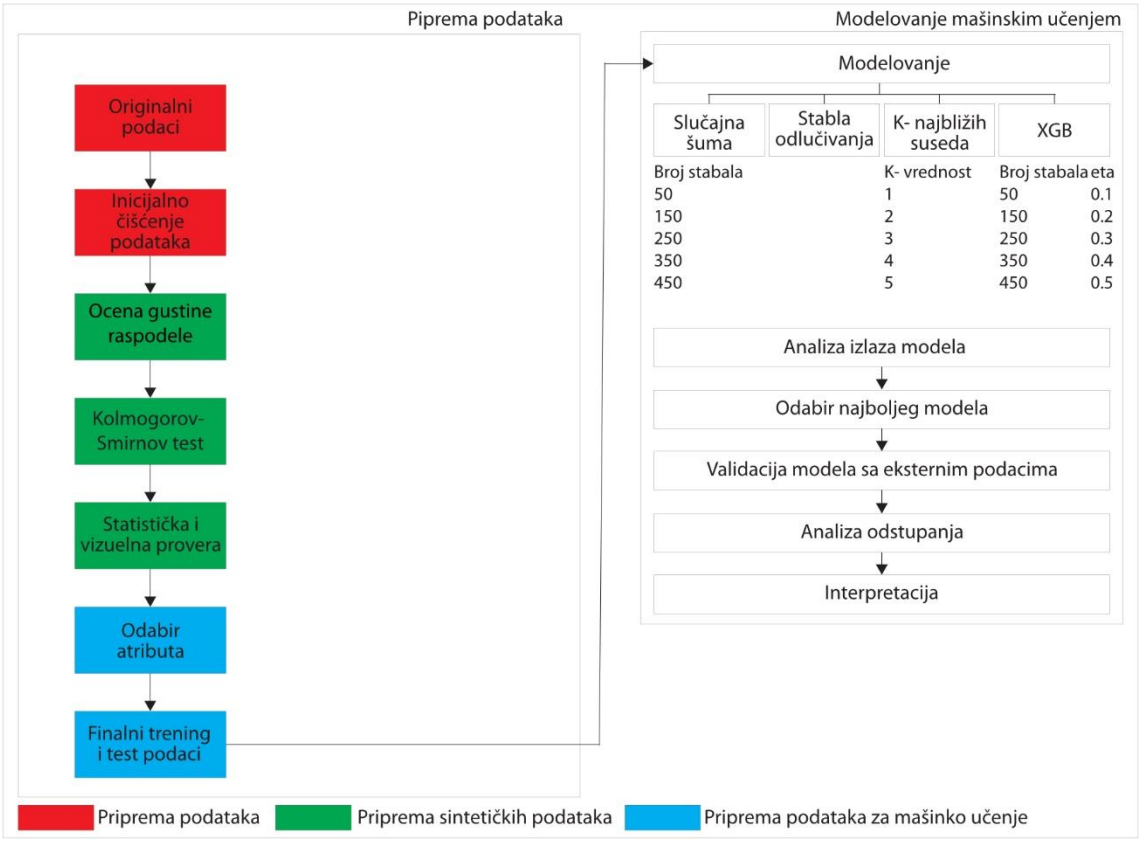
U ovom primeru biće prikazana primena mašinskog učenja za prognoziranje talasovodnih parametara oblasti D jonosfere. Ovaj primer je interesantan jer je oblast D jonosfere karakterisan sa dva parametra, pa će izlaz iz modela biti predstavljen sa dve ciljne promenljive. Takođe, zbog relativno malog uzorka podataka, biće primenjena metoda za progušćivanje seta podataka.

3.3.2.1. Korišćeni podaci, primenjene transformacije i obrada podataka

Metodologija istraživanja prikazana je na slici 27, gde je prikazano da su originalni podaci sadržavali ukupno 212 instanci, koje uključuju informacije o X- zračenju, razliku između amplitude i faze VLF signala, kao i talasovodne parametre niske jonosfere i dr. Dodatnih 45 uzoraka je isključeno iz ove analize i korišćeni su tek nakon odabira najboljeg modela za validaciju, tj. provere modela sa uzorcima koji nisu korišćeni prilikom treniranja i testiranja. Ocena gustine raspodele primenjena je za povećanje uzorka za potrebe mašinskog učenja.

Kao što je ranije pomenuto, minimalno potreban uzorak za mašinsko učenje nije detaljno definisan u literaturi. Potreba za većim brojem podataka zavisi od složenosti problema, kvaliteta atributa, odnosa između atributa i ciljne promenljive, kao i drugih faktora. Pošto nije bilo unapred poznato koliko uzoraka je potrebno, a pretpostavljajući da je 212 uzoraka premalo, primenjena je ocena gustine raspodele kako bi se procenila raspodela originalnih podataka, a potom je izvučeno ukupno 5000 sintetičkih podataka iz slične raspodele. Ovakav pristup treba primeniti samo ako je prikupljanje dodatnih uzoraka otežano, a alternativne metode ne postoje. Takođe, prilikom primene ovog metoda progušćivanja podataka, potrebno je biti oprezan pri

interpretaciji rezultata. U ovom istraživanju, oprez je bio prisutan naročito pri dobijanju progušćenih podataka, gde je za svaku promenljivu (atribut ili ciljnu promenljivu) primenjen Kolmogorov-Smirnov test kako bi se proverilo da li sintetički i originalni podaci dolaze iz iste raspodele, a potom je izvršena provera putem deskriptivne statistike i vizuelnog pregleda. Takođe, kao mera opreza, prethodno pomenutih 45 podataka je isključeno iz svih analiza i korišćeni su tek nakon odabira finalnog modela. Na kraju, model sa velikim brojem sintetičkih podataka ne može se smatrati finalnim modelom, ali rezultati dobijeni iz ovog modela mogu pružiti dobru indikaciju za pripremu većeg broja podataka, koji će služiti za procenu parametara oblasti D jonosfere.



Slika 27. Radni tok istraživanja sa pripremom podataka, pripremom sintetičkih podataka i modelovanjem mašinskim učenjem

Pored toga što je 45 podataka odvojeno na početku istraživanja za validaciju modela, set podataka za testiranje modela predstavljao je originalnih 212 uzoraka. Ovaj pristup treba primeniti sa oprezom, jer je korišćen samo za odabir najboljeg modela. Mere kvantifikacije modela regresije u tom slučaju ne koriste se kao finalne za set podataka za testiranje, već se koriste one iz seta podataka za validaciju. Ideja iza ovog pristupa zasniva se na prognoziranju

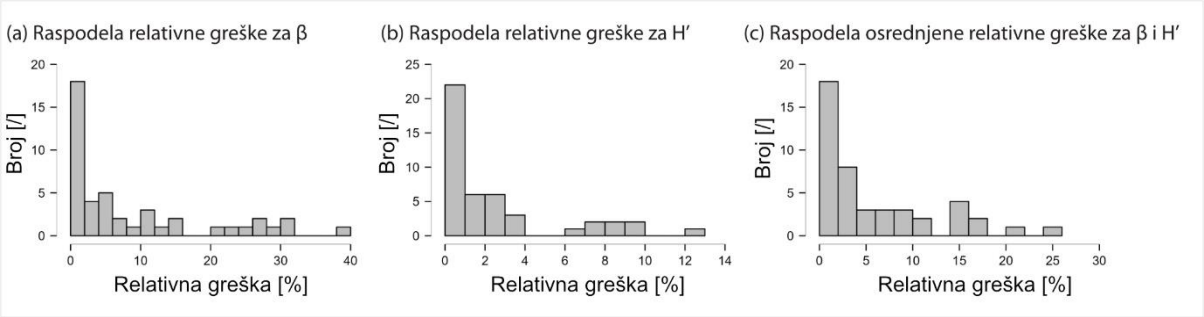
u uzorku i van uzorka, sa pretpostavkom da najbolji model za prognozu u uzorku jeste i najbolji za prognozu van uzorka. Ako model nije sposoban da postigne najbolje rezultate za podatke u uzorku, iz kojih je dobijen set podataka za treniranje, neće biti sposoban ni za najbolje rezultate na podacima koji se nisu nalazili u sintetičkom setu podataka. Ova dva postupka manipulacije podacima treba primeniti sa rezervom, i to samo u slučajevima kada nije lako doći do dodatnih podataka. U ovom istraživanju, oba postupka su primenjena, a u delu metodologije, kao i u interpretaciji rezultata, ovo je uzeto u obzir.

3.3.2.2. Rezultati aproksimacije talasovodnih parametara oblasti D jonosfere

Inicijalna analiza podataka nakon sintetičkog progušćenja pokazala je da Kolmogorov-Smirnov test nije ukazao na značajnu razliku između originalnih i sintetičkih uzoraka za svih pet instanci atributa i ciljnih promenljivih. Vizuelni test je takođe pokazao da su raspodele originalnih i sintetičkih podataka vizuelno slične, dok je deskriptivna statistika otkrila da originalna raspodela atributa X- zračenja pokazuje izrazitu asimetriju sa repom prema pozitivnim vrednostima. Kada su upoređeni koeficijenti asimetrije i spljoštenosti između originalnih podataka (6,178 i 42,932) i sintetičkih podataka (6,22 i 42,622), vrednosti su bile vrlo slične. Ove vrednosti, zajedno sa Kolmogorov-Smirnov testom i vizuelnom proverom, potvrđuju da su originalni i sintetički podaci uslovno isti. Postupak je ponovljen za sve atribute i ciljne promenljive.

U prvoj iteraciji modelovanja, modeli K- najbližih suseda i stabla odlučivanja su inicijalno odbačeni jer su prikazali veće vrednosti MAPE parametra za obe ciljne promenljive. Nakon toga, odluka između stabla odlučivanja i ekstremnog povećanja gradijenta ostala je ključna za odabir finalnog modela. Najbolji model slučajnih šuma bio je model sa 250 stabala, dok je najbolji model ekstremnog povećanja gradijenta imao 150 stabala i stopu učenja od 0,2. Model stabla odlučivanja prikazao je odstupanja u odnosu na stvarne vrednosti talasovodnih parametara oblasti D jonosfere od 1,061% i 0,017%, dok je model ekstremnog povećanja gradijenta imao vrednosti od 0,808% i 0,03%. Osrednje vrednosti relativne greške bile su relativno slične, ali poređenje maksimalne greške za parametar oštrine granice bilo je presudno. Model stabla odlučivanja prikazao je odstupanje od 11%, dok je model ekstremnog povećanja gradijenta imao vrednost od 3%. Važno je napomenuti da ove vrednosti odstupanja treba uzeti sa rezervom, jer su one dobijene sa setom podataka za testiranje u uzorku, te se očekuje da će vrednosti biti veće kada se model primeni na podatke koje nije imao tokom treniranja.

Slika 28 prikazuje raspodele relativne greške za obe ciljne promenljive za 45 primera iz validacionog seta podataka. Kao što je očekivano, MAPE parametar sračunat za primere iz validacionog seta podataka prikazuju veće, a ujedno i realnije vrednosti u opsegu od 9,1% za parametar oštrine granice i 2,45% za parametar visine granice reflektovanja. Takođe, za svaki primer, sračunato je i osrednjeno odstupanje između parametra oštrine i visine granice. Maksimalna odstupanja se nalaze u vrednosti od 38,8% za parametar oštrine granice i 12,2% za parametar visine granice reflektovanja. Ovakve vrednosti odstupanja su očekivane, pošto su uvećane od vrednosti koje su dobijene za prognoziranje u uzorku, ali prikazuju i relativno zadovoljavajuće i obećavajuće rezultate modela.

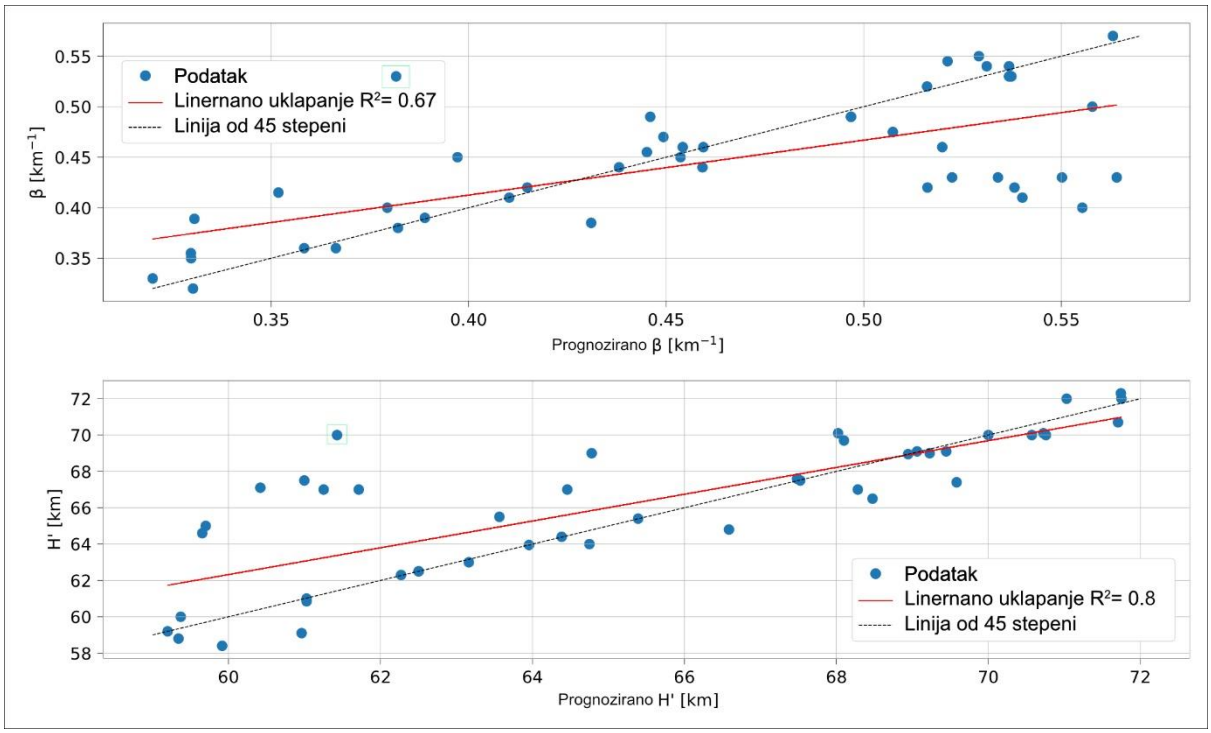


Slika 28. Odstupanja modela na validacionom setu podataka; (a) Parametar oštrine; (b) Parametar visine granice reflektovanja i (c) Osrednjene vrednosti parametra oštrine i visine granice reflektovanja

Potrebno je napomenuti da relativna greška za obe ciljne promenljive prikazuje izraženu asimetriju raspodele ka većim vrednostima. Detaljnijom analizom, može se uočiti da se 66% podataka za parametar oštrine nalazi ispod 10% greške, dok 55% podataka ima grešku manju od 5%. Sa druge strane, 97% podataka za parametar visine granice reflektovanja se nalazi ispod 10% relativne greške, dok 82% podataka ima relativnu grešku manju od 5%. Interpretacija ovih odstupanja može se svesti na nekoliko primera koji značajno unose asimetriju u raspodelu relativne greške. U većini slučajeva, model prikazuje relativno male prognozirane vrednosti, osim u nekoliko ekstremnih slučajeva.

Slika 29 prikazuje vrednosti prognoziranih i stvarnih vrednosti za parametre oštrine i visine granice reflektovanja VLF signala. Koeficijent determinacije za oba parametra je relativno zadovoljavajući, sa vrednostima od 0,67 za oštrinu i 0,8 za visinu granice reflektovanja. Linearno uklapanje podataka, kao i linija od 45 stepeni, pokazuju dobro poklapanje prognoziranih vrednosti sa stvarnim vrednostima, uz prostor za dalje unapređenje modela i podataka. Kao što se može videti na slici, postoje podaci sa većim odstupanjima (označeni zelenim kvadratom). U slučaju oštrine, prognozirana vrednost od 0,38 km^{-1} ima razliku od

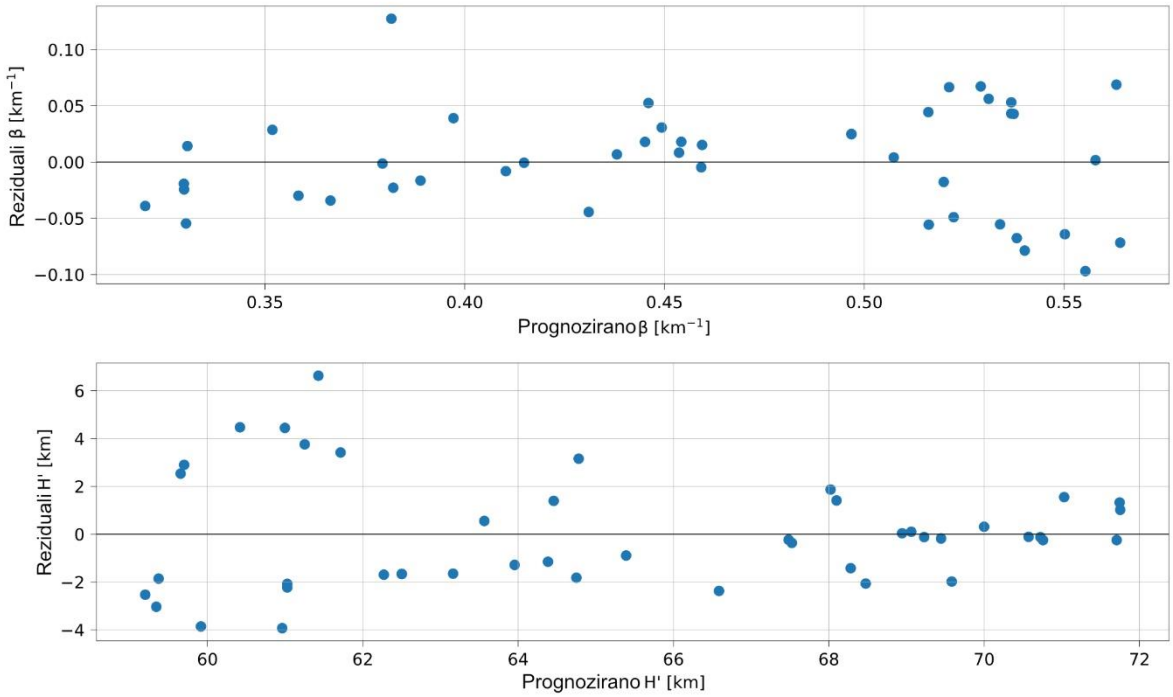
stvarne vrednosti od 0,53 km^{-1}, što daje relativnu grešku od 28%. Za parametar visine granice reflektovanja, isti prognozirani par pokazuje odstupanje od 0,82%, sa prognoziranom vrednošću od 70,6 km, dok je stvarna vrednost 70 km. Ovi podaci se odnose na solarni fler klase C4,8, koji nije visokog intenziteta, promenom amplitude i faze u odnosu na neporemećeni signal od 0,06 dB i fazi od 26 stepeni. Slična situacija nastaje kod solarnog flera C9,6, koji pokazuje razliku u amplitudi od 5,13 dB i fazi od 50,05 stepeni u odnosu na neporemećeni signal. U ovom slučaju, parametar oštrine je prognoziran sa relativnom greškom od 1,2% (prognozirana vrednost: 0,563 km^{-1}, stvarna vrednost: 0,57 km^{-1}), dok je parametar visine granice reflektovanja prognoziran sa većim odstupanjem (prognozirana vrednost: 61,4 km, stvarna vrednost: 70 km). Uzimajući u obzir početni kvaliteti podataka, može se reći da su rezultati zadovoljavajući i da pokazuju potrebu za unapređenjem ulaznih podataka.



Slika 29. Poređenje prognoziranih i istinitih vrednosti parametra oštrine i visine granice reflektovanja VLF signala

Dodatna analiza modela može se izvršiti računajući reziduale i grafički prikazivati prognozirane vrednosti i reziduale modela (Slika 30). Na slici 30, prilikom prikazivanja reziduala za parametar oštrine, ne može se primetiti nikakav šablon. Prognozirane vrednosti kreću se u opsegu od 0,05 do -0,05 km^{-1}, osim u nekoliko primera sa većim odstupanjima. Sa druge strane, kod reziduala za parametar visine granice mogu se uočiti određeni šabloni. Naime, pri prognozi viših vrednosti parametra visine granice model prikazuje manja

odstupanja, dok su odstupanja veća za niže vrednosti. Podelu je moguće postaviti na visini od 62 km, gde podaci sa prognozama iznad ove vrednosti imaju grešku u opsegu od -2 do +2 km, dok su podaci ispod 62 km povezani sa većim odstupanjima, od -4 do +6 km. U vrednostima relativne greške, podaci sa prognozama iznad 62 km imaju relativnu grešku od 1,23%, dok podaci ispod te granice imaju grešku od 5,15%. Interpretacija ovih rezultata može se povezati sa odnosom između X- klase solarnih flerova u podacima ispod 62 km. Naime, od 9 slučajeva X- klase solarnih flerova u validacionom setu podataka, 8 je imalo prognozirane vrednosti ispod 62 km. Ovo sugeriše da je modelu teže da prognozira jonosferske parametre u vrlo poremećenim uslovima, kao što su X- klasa solarnih flerova. Takođe, disbalans između broja X- klase solarnih flerova u trening setu (1,89%) i validacionom setu (20%) doprineo je povećanoj grešci pri prognoziranju jonosferskih parametara za X- klasu solarnih flerova.



Slika 30. Poređenje prognozirane vrednosti i reziduala za parametre oštrine i visine granice reflektovanja VLF signala

Važno je napomenuti da ovo nije loš rezultat jer je svrha validacionog seta podataka upravo da pokaže pristrasnosti modela prilikom prognoziranja određenih vrednosti, što je i slučaj ovde. Prognoziranje vrlo poremećenih vrednosti, poput X- klase solarnih flerova, je očekivana, a sličan problem se javlja i pri primeni LWPC softverskog paketa. Jedan od načina mitigacije ovog problema mogao bi biti prikupljanje većeg broja podataka za slučajeve X- klase solarnih flerova, koji su najređi. Ovim primerom je naglašena potreba za proširenim setom podataka.

3.4. Primena statističkih metoda na podatke magnetne susceptibilnosti iz uzoraka flotacijskog jalovišta rudnika „Rudnik“

U ovom delu biće prikazane statističke metode i njihova primena na podatke magnetne susceptibilnosti prikupljene sa flotacijskog jalovišta rudnika „Rudnik“, Republika Srbija. Glavna ideja uvođenja primera sa statističkim metodama, u odnosu na metode mašinskog učenja i prognoziranja vremenskih serija, jeste da se primenom različitih statističkih metoda dobiju dodatne informacije koje nisu dostupne standardno primenjenim metodama istraživanja. Pored tih metoda, primenjene su i metode koje su primarno vezane za podatke vremenskih serija (npr. stacionarnost) na podatke koji su predstavljeni kao prostorna serija. Oblast nauke o podacima obuhvata veliki broj metoda koje nisu sve povezane sa mašinskim učenjem i vremenskim serijama. Ovaj primer, zajedno sa ostalim primerima u disertaciji, predstavlja celokupan radni tok u kojem su prikazane različite metode primenjene na geofizičke, geološke i podatke atmosferske fizike, sa ciljem dobijanja dodatnih informacija o samim podacima.

Primer sa rudnika „Rudnik“ je interesantan iz više razloga. Naime, primena metode magnetne susceptibilnosti na materijal prikupljen sa jalovišta u cilju definisanja koncentracije teških metala predstavlja novitet. Takođe, količina podataka, sa obzirom na to da su uzorci uzimani sa međusobnim dubinskim rastojanjem od 10 cm, čini ovaj skup podataka visokog kvaliteta i visoke rezolucije. Sa druge strane, jalovišta i materijal prikupljen sa njih predstavljaju potencijalni ekološki hazard zbog prisustva teških metala, koji mogu imati negativne posledice po ljudski organizam ukoliko dospeju u njega (Liu et al., 2023). Takođe, ekološki problemi izazvani jalovištima mogu se kumulirati vremenom ili spontano, na primer usled zemljotresa ili poplava (Su et al., 2024). U Republici Srbiji, više ekoloških katastrofa je nastalo usled jalovišta, kao što su Valja Fundata (Majdanpek) 1974. godine, Šaški potok (Majdanpek) 1996. godine i Stolice (Krupanj) 2014. godine (Nišić et al., 2024). Sa druge strane, materijal sa jalovišta se u poslednje vreme smatra (geo)resursom (Cacciuttolo et al., 2023), a postoje slučajevi u kojima se materijal sa jalovišta reciklira za druge potrebe. Veliki broj savremenih istraživanja je pokazao da merenja magnetne susceptibilnosti u magnetnom polju niskog intenziteta mogu da omoguće određivanje zona sa povišenom koncentracijom teških metala, zato što fero(feri)magnetični minerali imaju afinitet prema teškim metalima (Bityukova et al., 1999; Boyko et al., 2004; Hanesch & Schloger, 2005; Kim et al., 2010; Salehi et al., 2013; Zawadzki et al., 2015; Brempong et al., 2016; Jaffar et al., 2017; Vasiliev et al., 2020). Budući

da se merenje magnetne susceptibilnosti uspešno primenjuje za određivanje prostorne raspodele teških metala u zemljištu (Lecoanet et al., 1999; Petrovský et al., 2000; Karimi et al., 2011; Wang, 2013; Brempong et al., 2016; Oudeika et al., 2020), kao logični nastavak bila je primena ove vrste merenja za određivanje dubinske raspodele teških metala. Primer takvih istraživanja su ispitivanja vertikalne distribucije magnetne susceptibilnosti na rudnim jalovištima (Jordanova et al., 2013, Gómez- García et al., 2015). Merenja magnetne susceptibilnosti za određivanje dubinske raspodele teških metala u jalovištima su veoma korisna jer omogućavaju da se dobiju podaci uz minimalnu pripremu uzoraka i na vrlo ekonomičan način, pri čemu uzorci ostaju neporemećeni za dalja mineraloška, geohemijska i druga ispitivanja.

Prilikom karakterizacije materijala iz flotacijskog jalovišta rudnika „Rudnik“, sa krajnjim ciljem ocene mogućnosti prerade materijala jalovišta za građevinsku industriju (Simić et al., 2024), zbog svojih pogodnosti, primenjena je metoda magnetne susceptibilnosti. Područje primene metode magnetne susceptibilnosti odnosilo se na karakterizaciju vertikalne raspodele teških metala u uzorcima materijala jalovišta.

U ovom podpoglavlju prikazana je primena nestandardnih statističkih metoda na podatke magnetne susceptibilnosti, kako bi se dobile dodatne informacije iz podataka koje nije moguće dobiti konvencionalnim metodama analize. Pored primene statističkih metoda, biće primenjene i metode za analizu vremenskih serija u cilju testiranja da li ove metode mogu da se primene na prostorno zavisne podatke i da li pružaju dodatne informacije.

3.4.1. Opis podataka i primenjenih metoda

Postavka ovog istraživanja u disertaciji je relativno jednostavna, primenjene su standardne metode deskriptivne statistike (srednja vrednost, medijalna vrednost, maksimum, minimum, koeficijent asimetrije i spljoštenosti, koeficijent varijacije, itd.). Takođe, upotrebljen je Hartiganov test modaliteta, koji proverava da li je raspodela unimodalna ili bimodalna, kao i prethodno prikazan Kolmogorov- Smirnov test. Analizirani su Pirsonov i Spirmanov koeficijent korelacije, kao i metode koje se primenjuju u analizi vremenskih serija- testovi stacionarnosti. Pored tih metoda vremenskih serija, primenjeni su i grafici kumulativne sume, kao i grafici prve diference.

Prva iteracija istraživanja bila je fokusirana na primenu statističkih metoda za međusobno poređenje tri bušotine kako bi se dobile informacije o njihovoj sličnosti ili različitosti. Ideja iza ovog koraka bila je da se primene metode korelacije, a pored njih, i Kolmogorov- Smirnov test, kako bi se odredilo da li materijal iz jedne bušotine dolazi iz iste raspodele kao materijal iz druge bušotine. Takođe, za istu upotrebu primenjene su i standardne statističke metode.

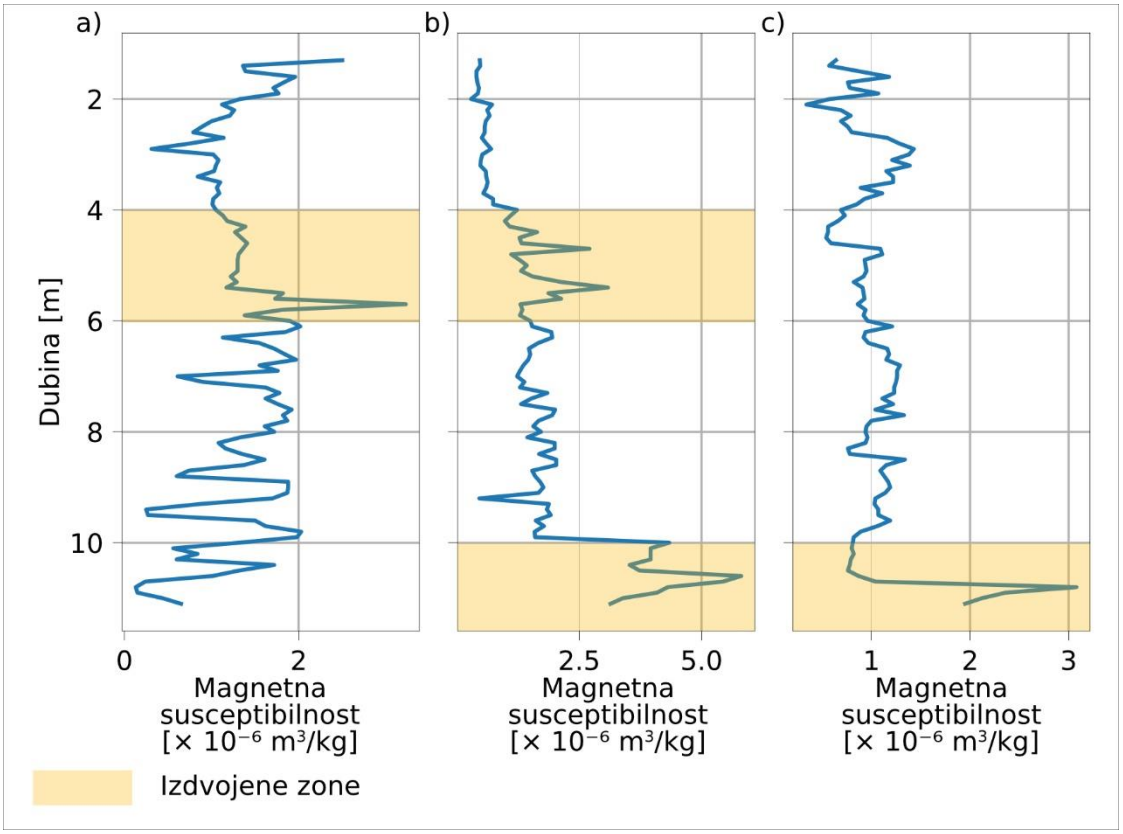
Druga iteracija istraživanja predstavljala je primenu istih statističkih metoda za dobijanje informacija i statističke značajnosti o zonama od interesa u bušotinama. Zone od interesa definisane su kao slojevi sa povišenom koncentracijom teških metala, odnosno povišenim vrednostima magnetne susceptibilnosti. Prilikom izdvajanja slojeva od interesa najpre je korišćena vizuelna metoda, a nakon toga su se izvodili statistički testovi za određivanje da li postoji statistička značajnost u odstupanju tog sloja u odnosu na slojeve neposredno ispod ili iznad datog sloja. Istraživanje je sprovedeno upotrebom samo podataka magnetne susceptibilnosti, bez informacija o litologiji materijala iz bušotine. Time je omogućeno da se rezultati istraživanja uporede sa litološkim stubom bušotine, čime se odredila validnost informacija dobijenih iz istraživanja.

Metode analize vremenskih serija (stacionarnost, grafici prve diference podataka i kumulativne sume) primenjene su za testiranje da li ove metode mogu pružiti dodatne informacije prilikom analiziranja prostorno zavisnih podataka. Glavna pretpostavka pri analizi vremenskih serija jeste da vremenske serije trebaju biti uzorkovane u vremenski istim intervalima. U ovom primeru, ovo je omogućeno za prostorno zavisne podatke, tj. da su oni uzorkovani u prostorno istim intervalima (10 cm).

3.4.2. Rezultati primene statističkih metoda na podatke magnetne susceptibilnosti

Slika 31 prikazuje raspodelu magnetne susceptibilnosti sa dubinom materijala jalovišta rudnika „Rudnik“. Vizuelnom analizom nije moguće detaljno odrediti sličnosti između tri međusobne bušotine. Izdvojene zone, zasnovane na vizuelnom pregledu bušotina, prikazuju lokalne koncentracije povišenih vrednosti magnetne susceptibilnosti. Ukoliko se te tri bušotine međusobno uporede, postoji donekle neka sličnost. Na primer, bušotina RJ- 2 prikazuje na sličnim lokacijama povišene vrednosti magnetne susceptibilnosti- zona od četvrtog do šestog metara zajedno sa bušotinom RJ- 1, kao i zona na 10+ metara zajedno sa bušotinom RJ- 3.

Bušotina RJ- 1 prikazuje opadajuće vrednosti magnetne susceptibilnosti od površi terena do četvrtog metra, dok od četvrtog do šestog metra prikazuje blagi porast sa jednim izraženim maksimumom pre šestog metra. Nakon šestog metra, vrednosti magnetne susceptibilnosti naglo variraju od većih ka nižim vrednostima. Bušotina RJ- 2 prikazuje relativno stabilne vrednosti do četvrtog metra, nakon čega, slično kao kod bušotine RJ- 1, mogu se uočiti povećane vrednosti. Samo u ovom slučaju postoje dva lokalna maksimuma magnetne susceptibilnosti. Nakon šestog metra, vrednosti su vrlo stabilne do desetog metra, nakon čega se može videti izraženo povećanje vrednosti magnetne susceptibilnosti. Bušotina RJ- 3 prikazuje vrlo umerene vrednosti magnetne susceptibilnosti od površine terena do desetog metra, gde se tek od desetog metra može uočiti povećanje vrednosti magnetne susceptibilnosti.



Slika 31. Magnetna susceptibilnost na tri bušotine sa jalovišta rudnika „Rudnik“ (a) RJ- 1; (b) RJ- 2 i (c) RJ- 3

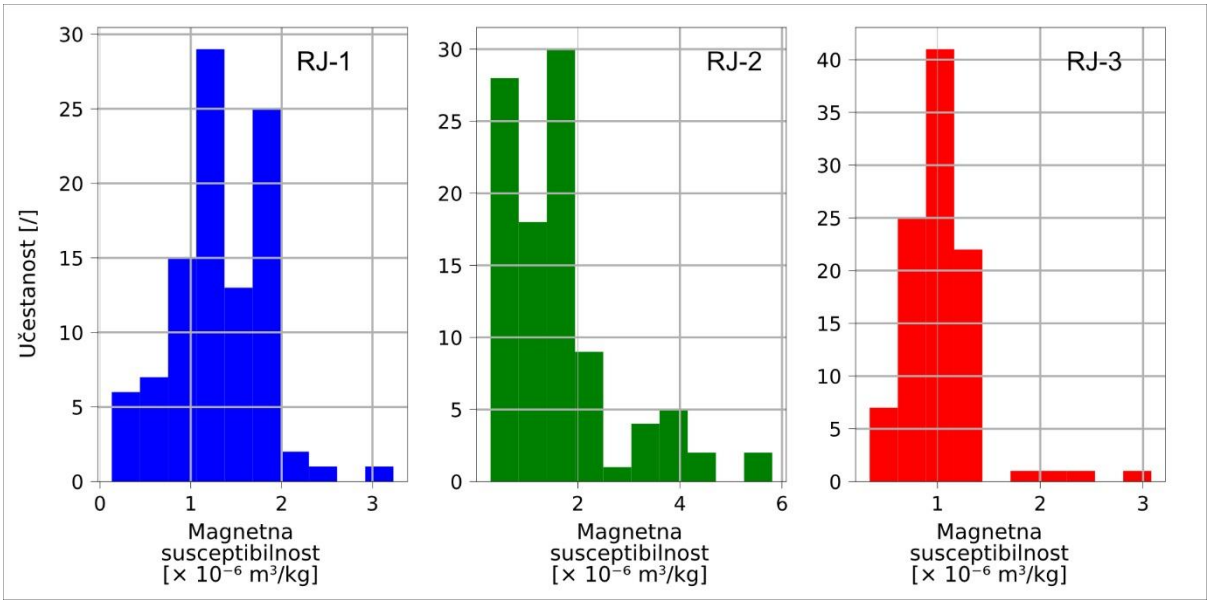
Vizuelna analiza nije pružila nikakve informacije o međusobnoj sličnosti između materijala iz tri bušotine, osim sličnih zona povećane magnetne susceptibilnosti, čija će statistička validnost biti proverena naknadno. Naredni korak predstavlja analizu standardne deskriptivne statistike sračunate za vrednosti magnetne susceptibilnosti za sve tri bušotine. Poređenjem koeficijenta

varijacije između tri bušotine može se videti da bušotina RJ- 2 prikazuje veće vrednosti za 29% u odnosu na bušotinu RJ- 1 i za 34% u odnosu na bušotinu RJ- 3, pri čemu su bušotine RJ- 1 i RJ- 3 međusobno slične. Parametar maksimalne vrednosti prikazuje sličnu situaciju kao parametar koeficijenta varijacije, gde su bušotine RJ- 1 i RJ- 3 međusobno uporedive, dok bušotina RJ- 2 prikazuje više vrednosti. Slična situacija se javlja i sa varijansom i opsegom.

Tabela 8. Parametri deskriptivne statistike za bušotine RJ-1, RJ-2 i RJ-3

Parametar/ Bušotina	RJ-1	RJ-2	RJ-3
Srednja vrednost [× 10^{-6} m^3/kg]	1.32	1.63	1.03
Medijalna vrednost [× 10^{-6} m^3/kg]	1.31	1.47	0.96
Minimum [× 10^{-6} m^3/kg]	0.13	0.29	0.34
Maksimum [× 10^{-6} m^3/kg]	3.23	5.81	3.08
Opseg [× 10^{-6} m^3/kg]	3.10	5.52	2.74
Varijansa [× 10^{-6} m^3/kg]	0.27	1.26	0.13
Koeficijent varijacije [/]	0.40	0.69	0.35
Koeficijent iskrivljenosti [/]	0.11	1.55	2.52
Koeficijent ispupčenosti [/]	0.98	2.47	10.73

Prilikom analize raspodele magnetne susceptibilnosti materijala iz tri bušotine, vizuelno se uočava da raspodele RJ- 1 i RJ- 2 potencijalno prikazuju bimodalnost, dok je RJ- 3 očigledno unimodalna (Slika 32). Radi provere unimodaliteta ili bimodaliteta, primenjen je Hartiganov test za modalitet. Hartiganov test je pokazao da su bušotine RJ- 1 i RJ- 3 unimodalne (RJ- 3 je testirana iako je bilo očigledno da ne prikazuje naznake bimodaliteta), dok je RJ- 2 bimodalna. Test je još jednom pokazao različitosti između bušotina RJ- 1, RJ- 3 i RJ- 2. Moguća interpretacija bimodalnosti bušotine RJ- 2 može se povezati sa postojanjem dve izdvojene zone povećane magnetne susceptibilnosti.



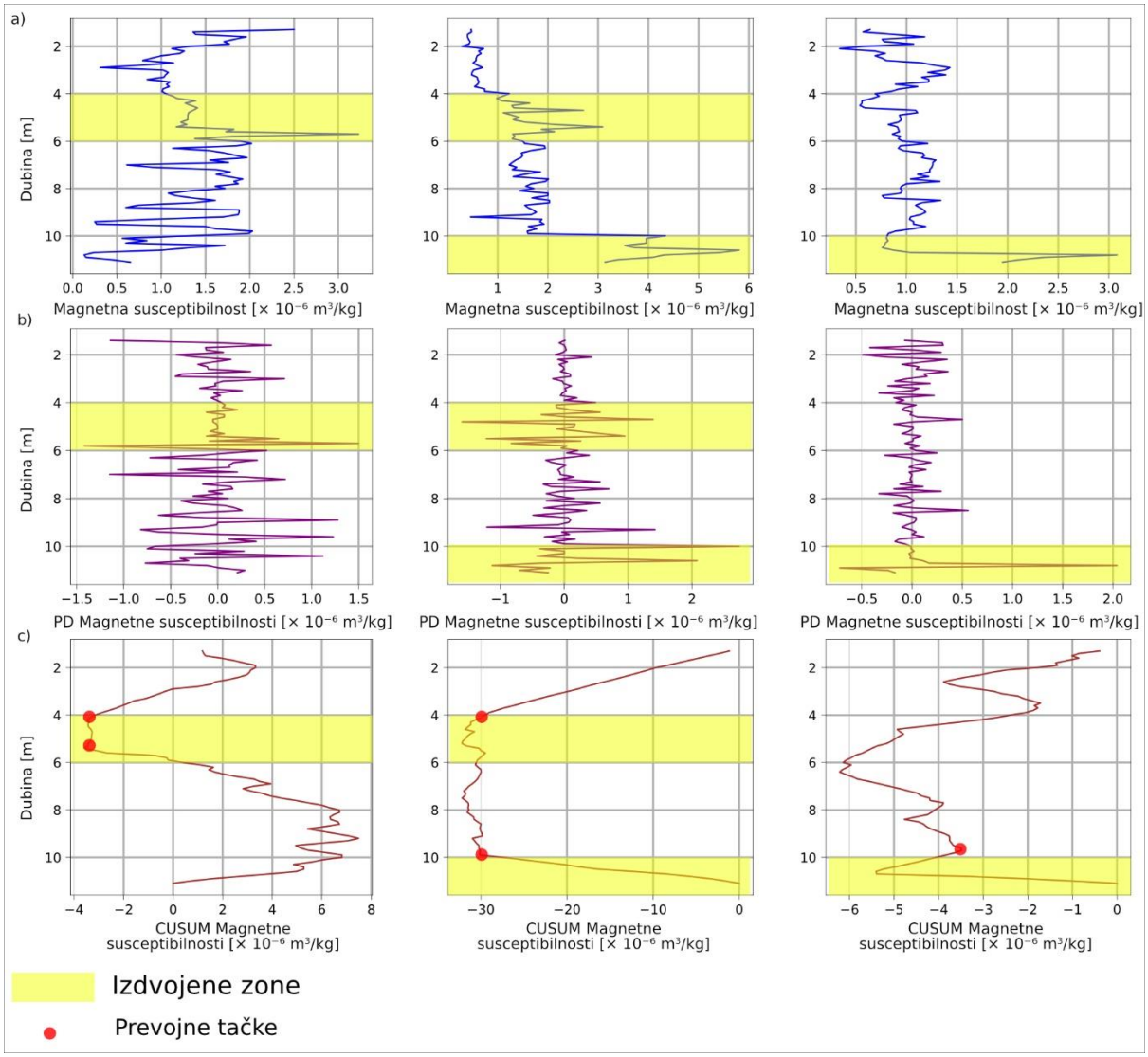
Slika 32. Raspodele podatka magnetne susceptibilnosti za bušotine RJ-1, RJ-2 i RJ-3

Primena Kolmogorov-Smirnov testa pokazala je da bušotine RJ- 1 i RJ- 2 prikazuju sličnost (KS statistika 0,17, p- vrednost $1{,}08 \times 10^{-1}$), dok bušotine RJ- 1 i RJ- 3 (KS statistika 0,45, p- vrednost $1{,}36 \times 10^{-9}$) i RJ- 2 i RJ- 3 (KS statistika 0,55, p- vrednost $7{,}26 \times 10^{-14}$) prikazuju međusobnu različitost. Ukoliko se samo porede rezultati Kolmogorov-Smirnov testa, može se zaključiti da su bušotine RJ- 1 i RJ- 2 slične, dok su bušotine RJ- 1 i RJ- 3, kao i RJ- 2 i RJ- 3, međusobno različite. Sa druge strane, prethodno prikazani rezultati vizuelne analize, deskriptivne statistike i raspodele podataka pokazuju drugačiju sliku. Analiza koeficijenata korelacije pokazala je da bušotine međusobno imaju vrlo malu korelaciju, pri čemu su korelacije između RJ- 1 i RJ- 2 -0,29 (Pirson) i -0,063 (Spirman), između RJ- 1 i RJ- 3 -0,32 (Pirson) i -0,061 (Spirman), i između RJ- 2 i RJ- 3 0,28 (Pirson) i 0,11 (Spirman). Svi koeficijenti korelacije su zanemarljivi prema raspodeli opsega koeficijenata korelacije prikazanoj u Schober et al. (2018). Primenjeni testovi stacionarnosti takođe ukazuju na međusobnu različitost bušotina RJ- 1 i RJ- 3 u odnosu na RJ- 2, pri čemu su bušotine RJ- 1 i RJ- 3 stacionarne prema proširenom Diki-Fulerovom testu (ADF statistika -5,03 i -3,71, kritična vrednost -2,89) i prema Kviatkovski- Filipps- Šmit- Šin testu (KPSS statistika 0,25 i 0,43, kritična vrednost 0,463), dok je bušotina RJ- 2 nestacionarna prema oba prethodna testa (ADF statistika -1,62 i KPSS statistika 1,16).

Interpretacija prvog dela rezultata je vrlo otežana. Rezultati koji su dobijeni ukazuju na potpunu međusobnu različitost između sve tri bušotine, dok Kolmogorov- Smirnov test pokazuje da

postoji sličnost između bušotina RJ- 1 i RJ- 2. Interpretacija rezultata stacionarnosti može se povezati sa tim da stacionarne bušotine (RJ- 1 i RJ- 3) prikazuju uniformno deponovanje materijala, dok je nestacionarna bušotina (RJ- 2) pokazala neuniformnost u deponovanom materijalu. Takvu pretpostavku podržavaju izraženo veći koeficijent varijacije, kao i postojanje dve zone lokalno povišenih vrednosti magnetne susceptibilnosti. Otežavajuća okolnost prilikom određivanja sličnosti u materijalu između ove tri bušotine dodatno je otežana time što je materijal sa jalovišta antropogenog porekla. Geološki, taj materijal je poreklom lokalan, ali je nakon obrade za izdvajanje korisnog materijala izmešan i deponovan relativno nasumično.

Slika 33 prikazuje originalne podatke magnetne susceptibilnosti, prvu diferencu magnetne susceptibilnosti i kumulativnu sumu magnetne susceptibilnosti. Na osnovu originalnih podataka izdvojene su ukupno četiri zone lokalnih povećanja magnetne susceptibilnosti, gde se dve nalaze na bušotini RJ- 2, a po jedna na bušotinama RJ- 1 i RJ- 3. Grafik prve diference podataka magnetne susceptibilnosti prikazuje povećane vrednosti, odnosno veću varijaciju unutar podataka u zonama koje su izdvojene. Na bušotini RJ- 1 situacija nije toliko očigledna, dok je na bušotini RJ- 2 i RJ- 3 situacija jasnija. Podaci kumulativne sume prikazuju zanimljive podatke. Najpre, za bušotinu RJ- 1 može se videti postojanje dve prevojne tačke: prva se nalazi na četvrtom metru, a druga malo iznad šestog metra. Podaci pre četvrtog metra prikazuju generalan silazni trend, dok podaci nakon šestog metra pokazuju uzlazni trend do devetog metra, nakon čega trend opet postaje silazni. Bušotina RJ- 2 prikazuje silazni trend do četvrtog metra, nakon čega je trend generalno stabilan do desetog metra, gde se može videti uzlazni trend. Bušotina RJ- 3 prikazuje više promena trenda podataka, pri čemu se jedna promena poklapa sa izdvojenom zonom.



Slika 33. (a) Podaci magnetne susceptibilnosti; (b) Prvi diferencijal podataka magnetne susceptibilnosti i (c) Kumulativna suma podataka magnetne susceptibilnosti

Poređenje izdvojenih zona sa zonama neposredno ispod (ukoliko postoji prostora za sloj ispod) ili iznad je prikazano na tabeli 9. Izdvojeni sloj u bušotini RJ- 1 prikazuje prema Kolmogorov-Smirnov testu razliku od sloja ispod (6- 8 metara) i iznad (2- 4 metara), dok izdvojeni sloj u bušotini RJ- 2 prikazuje izraženu razliku od sloja iznad datog sloja (2- 4 metara) dok ne prikazuje nikakvu razliku od sloja ispod (6- 8 metara). Kod bušotine RJ- 3 izdvojena zona ne prikazuje različitost od sloja koji joj prethodi (8- 10 metara).

Tabela 9. Rezultati primene Kolmogorov- Smirnov testa za proveravanje različitosti susednih slojeva

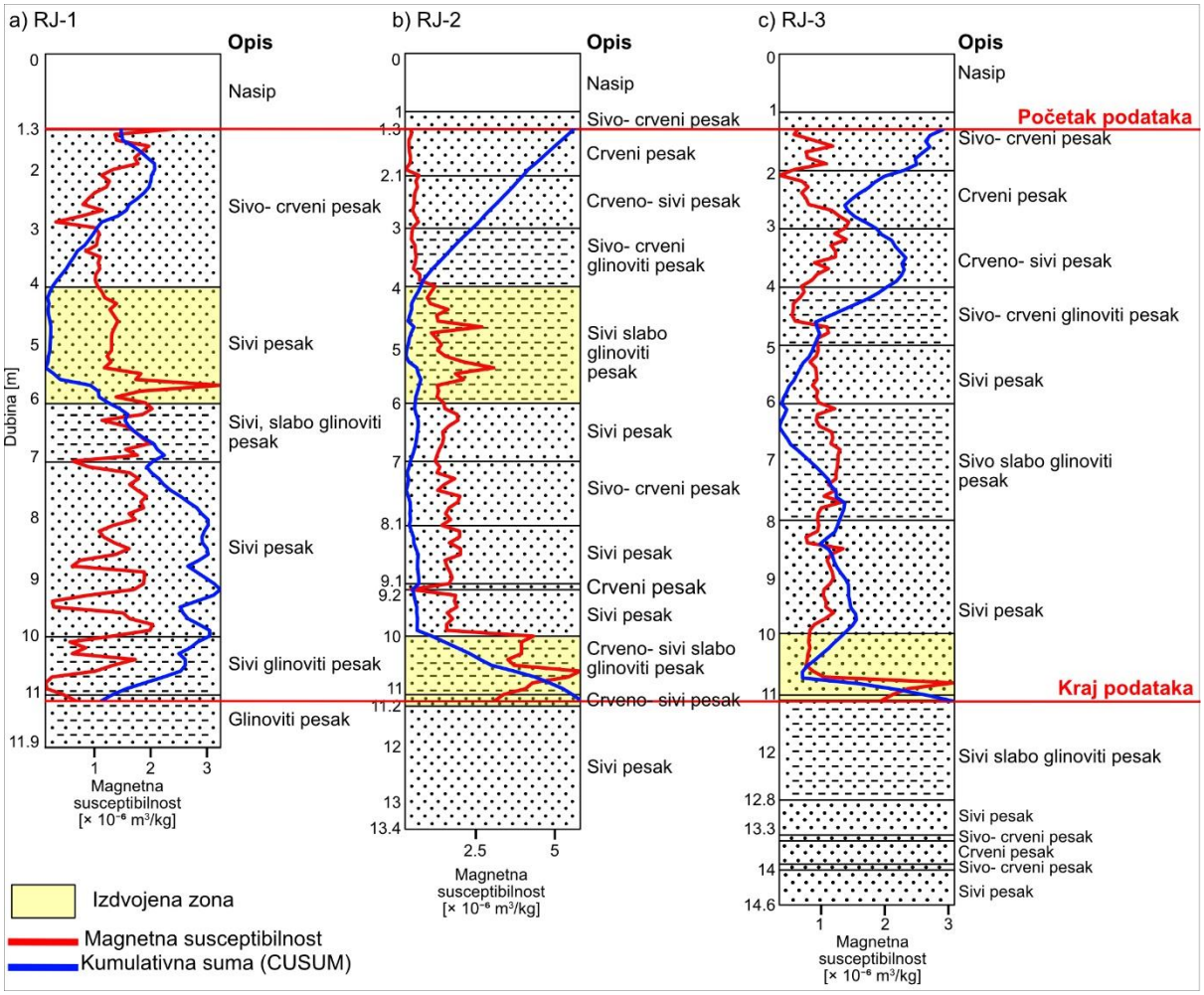
Bušotina	Zona	KS statistika	KS p- vrednost
		Sloj pre	
RJ-1	**1**	0,75	$9{,}55 \times 10^{-6}$
RJ-2	**1**	1,00	$1{,}45 \times 10^{-11}$
RJ-2	**2**	0,60	$1{,}12 \times 10^{-3}$
RJ-3	**1**	0,35	$1{,}75 \times 10^{-1}$
		Sloj nakon	
RJ-1	**1**	0,65	$2{,}7 \times 10^{-4}$
RJ-2	**1**	0,4	$8{,}11 \times 10^{-2}$

Slika 34 prikazuje korelaciju sa litološkim podacima. Kao što je prethodno pomenuto, korelacija sa litološkim podacima urađena je tek nakon dobijanja informacija o izdvojenim zonama i međusobnoj sličnosti između bušotina. Sličnost između bušotina je konstatovana kao vrlo mala, te se samo određene zone između bušotina mogu smatrati sličnim, dok u celosti bušotine ne prikazuju sličnost. Interpretacija ovog nalazi se u činjenici da je deponovanje materijala bilo antropogeno kontrolisano, odnosno da je materijal lokalnog porekla, ali je tokom prerade rude materijal deponovan nasumično. Korelacija izdvojenih zona sa slojevima pre i nakon datog sloja prikazala je da je kod bušotine RJ- 1 izdvojeni sloj zaseban sa statističkom značajnošću u odnosu na sloj pre i nakon datog sloja. Bušotina RJ- 2 prikazuje da je prvi izdvojeni sloj različit od sloja pre, ali nije različit od sloja nakon, dok je druga izdvojena zona različita od sloja koji joj prethodi. Bušotina RJ- 3 za svoju jednu izdvojenu zonu prikazala je da se statistički ne razlikuje od sloja pre nje.

Korelacija sa litološkim podacima za bušotinu RJ- 1 prikazuje da se ceo izdvojeni sloj u potpunosti poklapa sa litološkim slojem sivih peskova. Sloj koji prethodi izdvojenom sloju je sloj sivo-crvenih peskova, dok nakon njega dolaze sivi, slabo glinoviti peskovi. Grafik kumulativne sume prikazuje prevojnu tačku tačno na četvrtom metru, odnosno na početku sloja sivih peskova, gde trend menja iz silaznog u relativno konstantan. Druga prevojna tačka se nalazi unutar sloja sivih peskova nakon petog metra, gde se prikazuje uzlazni trend. Pri dnu sloja sivih peskova može se videti i maksimum za ceo set podataka. Izdvojena zona je testirana Kolmogorov-Smirnov testom, koji je pokazao da se razlikuje od sloja pre i sloja nakon izdvojene zone.

Korelacija litoloških podataka sa bušotinom RJ- 2 pokazuje da se prevojna tačka nalazi u sloju sivih, slabo glinovitih peskova koji čine izdvojenu zonu na osnovu samo podataka magnetne susceptibilnosti. U tom sloju se menja generalni trend podataka kumulativne sume, koji ostaje do desetog metra. Dva lokalna maksimuma se nalaze u izdvojenom sloju sivih, slabo glinovitih peskova. Kao i u slučaju sa bušotinom RJ- 1, kod bušotine RJ- 2 izdvojena zona u potpunosti odgovara sloju sivih, slabo glinovitih peskova. Na desetom metru u bušotini RJ- 2 prevojna tačka se poklapa sa početkom sloja crveno-sivih slabo glinovitih peskova, kojima prethodi sloj sivih peskova. Za prvu zonu kod bušotine RJ- 2 prikazano je da se razlikuje od sloja koji joj prethodi, ali ne i od sloja koji sledi. Interpretacija može biti vezana za grafik kumulativne sume, odnosno na početku četvrtog metra nalazi se prevojna tačka gde se menja trend kumulativne sume sa silaznog na relativno konstantan. U slučaju bušotine RJ- 1, izdvojena zona ima dve prevojne tačke, a grafik kumulativne sume je konstantan kroz celu zonu. Kod bušotine RJ- 2, grafik kumulativne sume ostaje konstantan i nakon izdvojene zone, što može biti razlog zašto se izdvojeni sloj razlikuje od sloja pre, ali ne i od sloja nakon. Kod druge izdvojene zone postoji razlika između izdvojene zone i sloja koji joj prethodi, a takođe u tom slučaju postoji prevojna tačka na granici litološkog stuba, odnosno na desetom metru.

Bušotina RJ- 3 prikazuje jednu izdvojenu zonu koja odgovara sloju sivih peskova. Zona koja je izdvojena počinje od desetog metra, dok sam sloj sivih peskova počinje od osmog metra. Za ovu zonu nije bilo potvrde od strane Kolmogorov- Smirnov testa da se razlikuje od sloja koji joj prethodi, a sa litološkog stuba se može videti da je to jedan sloj se prostire od osmog do jedanaestog metra.



Slika 34. Korelacija podataka magnetne susceptibilnosti i kumulativne sume magnetne susceptibilnosti sa litološkim podacima sa tri bušotine sa jalovišta rudnika „Rudnik“

4. Diskusija i buduća istraživanja

Diskusija dobijenih rezultata i predlozi za buduća istraživanja prikazani su u ovom poglavlju. Slično poglavlju 3, poglavlje 4 je grupisano i razvrstano prema oblastima istraživanja kojima pripadaju.

Primeri predstavljeni u disertaciji obuhvataju različite oblasti, uključujući geofiziku, fiziku atmosfere i fiziku jonosfere. Iako se razlikuju po specifičnostima, oblastima primene, prikazanim informacijama i upotrebi, njihova zajednička karakteristika u okviru ove disertacije jeste primena modela vođenih podacima na podatke iz tih oblasti. Modeli vođeni podacima omogućavaju dobijanje dodatnih informacija iz postojećih podataka, automatizaciju dugotrajnih i zahtevnih istraživačkih procesa ili predstavljaju alternativu postojećim metodama dobijanja određenih parametara kao što je prikazano u rezultatima ove disertacije.

4.1. Prognoziranje i imputacija koncentracija zagađujućih materija u vazduhu

U poglavlju koje se odnosi na rezultate vezane za koncentraciju zagađujućih materija prikazana su dva primera: prvo, primer primene Fejsbukovog Profet modela za prognozu budućih, neizmerenih vrednosti koncentracija $PM_{2.5}$ i PM_{10} u vazduhu (3.1.1.), a zatim primer primene modela slučajne šume za imputaciju podataka $PM_{2.5}$ na više lokaliteta u Republici Srbiji (3.1.2.).

Oba primera u ovom poglavlju bave se sličnim pitanjem, a to je aproksimacija neizmerenih podataka koncentracije zagađujućih materija u vazduhu. U prvom primeru, aproksimacija je predstavljena prognoziranjem budućih vrednosti koncentracije zagađujućih materija u vazduhu, dok drugi primer prikazuje aproksimaciju preskočenih opservacija.

Rezultati prognoze vremenskih serija pokazali su zadovoljavajuće rezultate, jer je primenjeni model uspešno prikazao vrlo niska odstupanja prognoziranih vrednosti u odnosu na stvarne vrednosti, ali su takođe zabeležena i veća odstupanja u pojedinim slučajevima. Slična situacija uočena je i u drugom primeru. Visoka odstupanja prognoziranih vrednosti u odnosu na stvarne vrednosti u slučaju podataka o koncentraciji zagađujućih materija u vazduhu mogu se pripisati

neočekivano visokim nivoima zagađenja. U tim situacijama, ukoliko model nije obučen na velikom skupu podataka, može mu nedostajati dovoljno informacija o ponašanju varijacija koncentracije zagađujućih materija. Jedan od mogućih načina da se ovaj problem izbegne jeste korišćenje skupa podataka koji pokriva duže vremenske periode, kako bi se obuhvatile varijacije koje nastaju usled nižih temperatura i upotrebe materijala za grejanje domaćinstava.

Fejsbukov Profet model je pokazao svoje prednosti u pogledu jednostavnosti primene i lakoće dobijanja prognoza. Sa druge strane, moguće je primeniti Fejsbukov Profet model kao deo većeg modela koji bi predstavljao hibridni pristup, pri čemu bi izlaz Profet modela bio korišćen kao atribut u novom modelu. Buduća istraživanja biće usmerena na primenu hibridnog pristupa u kombinaciji sa metodama mašinskog učenja kako bi se poboljšalo dobijanje vrednosti koncentracije zagađujućih materija u vazduhu.

Primer koji se ticao primene dvosmerne imputacije podataka metodama mašinskog učenja nije pokazao da predstavlja bolje rešenje u odnosu na jednostavne modele. Ideja celokupnog rada bila je da se razvijenom algoritmu obezbede najbolji mogući uslovi kako bi se postigli rezultati koji nadmašuju performanse jednostavnih metoda. Prethodno pomenuti najbolji uslovi uključuju što veću dužinu trajanja uvodne vremenske serije koncentracije zagađujućih materija, poređenje sa jednostavnim metodama poput imputacije srednjom vrednošću i medijanom, kao i primenu relativno malih vrednosti preskočenih opservacija (24, 48 i 72 sata). Rezultati su pokazali da je model ostvario uporedive vrednosti u odnosu na jednostavne metode, iako je njegova računarska potrošnja bila znatno veća. Na primer, u proseku je za svih devet stanica monitoringa koncentracije zagađujućih materija bilo potrebno oko 13 minuta za predloženi algoritam, dok je za jednostavne metode to bilo gotovo instantno. Uporedivi rezultati predloženog algoritma u poređenju sa jednostavnim metodama, uz značajno povećano računarsko vreme, ne opravdavaju njegovo korišćenje.

Takođe, potrebno je napomenuti da je prikazan primer vrlo idealizovan u kontekstu dobijanja podataka. U ovom primeru korišćeni su podaci koji variraju u opsegu od 8735 kontinuirano merenih sati do 4715 sati. U stvarnim uslovima, podaci o koncentraciji zagađujućih materija mogu sadržati vrlo visoku učestalost kratkoročnih preskočenih opservacija i nisku učestalost dugoročnih preskočenih opservacija. Drugim rečima, teško je pronaći idealan slučaj kao što je ovaj, gde postoji velika količina kontinuirano merenih podataka. Takođe, algoritam je pokazao da je odabir atributa ključan prilikom modelovanja, jer se u ovom primeru koristi samo jedan

parametar, umesto različitih grupa atributa, što predstavlja dodatni hiperparametar i povećava kompleksnost algoritma, a time i uvećava utrošeno računarsko vreme.

Pored prethodno navedenih ograničenja, postoji i ograničenje u samom algoritmu koje se odnosi na generisanje jedne vrednosti imputacije koja potom ulazi u set atributa za generisanje naredne vrednosti imputacije. Ovaj pristup zavisi od „tačnosti" svake prethodne generisane vrednosti imputacije, pa ukoliko se uvede veliko odstupanje u jednu imputiranu vrednost, svi naredni rezultati će prenositi istu grešku u daljim prognozama. Ovaj problem se može izbeći primenom generisanja više koraka imputacije u jednom trenutku, što je planirano za realizaciju u budućim istraživanjima.

Potrebno je napomenuti da je gotovo nemoguće osmisliti i razviti metodu koja će vršiti imputaciju ili prognozu za sve slučajeve. Koncentracija zagađujućih materija zavisi u velikoj meri od niza parametara, kao što su meteorološki uslovi, saobraćaj, lokalna geografija, vremenski period u godini, nedeljni periodi, mesec i drugi parametri čiji uticaj na koncentraciju zagađujućih materija u vazduhu još uvek nije u potpunosti istražen. Jedan od tih parametara je visina sloja planetarne granice (eng. *Planetary Boundary Layer- PBL*). Parametar PBL može imati značajnu ulogu u prognoziranju i imputaciji podataka koncentracije zagađujućih materija, jer tokom noći i u hladnijim mesecima sloj PBL se nalazi niže u odnosu na dnevne vrednosti ili toplije mesecima. Sa druge strane, primena drugih metoda za dvosmernu imputaciju podataka takođe može pružiti kvalitetne rezultate, ali je potrebno sprovođenje daljih istraživanja kako bi se odredilo da li je ovaj pristup zaista adekvatan.

Pored toga što je u ovoj disertaciji u oba primera prikazano modelovanje koncentracije zagađujućih materija u vazduhu u vidu vremenskih serija, potrebno je napomenuti da je koncentracija zagađujućih materija u vazduhu prostorno-vremenska serija, odnosno da se oba aspekta- prostorni i vremenski- moraju analizirati. Dalje istraživanje, osim prethodno pomenutih potencijalnih pravaca, uključiće i prostornu komponentu koncentracije zagađujućih materija u vazduhu, čime će se dobiti šira slika varijacije koncentracije zagađujućih materija.

Područje istraživanja koje se odnosi na koncentraciju zagađujućih materija u vazduhu vrlo je atraktivno i verovatno će ostati ili čak povećati svoju važnost u narednom periodu. Kvalitet vazduha koji se udiše predstavlja značajan javno-zdravstveni parametar za celokupnu populaciju, a ne samo za određene ugrožene grupe. Zbog toga, nastavak istraživanja o kvalitetu

vazduha u Republici Srbiji, prostorno-vremenskim uslovima i karakteristikama, kao i doprinosu različitih privrednih grana i ljudskih aktivnosti povećanom zagađenju vazduha, predstavlja vrlo bitno istraživačko područje.

4.2. Prostorna klasifikacija ofiolita istočne Vardarske zone

Prostorna klasifikacija litologija prikazala je zanimljive i korisne rezultate u pogledu interpretacije i statističke značajnosti, kao i u smislu implikacija za buduća istraživanja. Rezultati prve iteracije pokazali su F1- meru za klasu ofiolita od 0,17, dok je ta vrednost kasnije povećana na 0,31 promenom modela slučajne šume u ekstremno povećanje gradijenta i dodavanjem atributa udaljenosti od reka. Iako su rezultati pokazali mogućnost unapređenja, apsolutna vrednost F1- mere je mala, ali se može primetiti da je model uspešno detektovao prostorni položaj ofiolita u istočnoj Vardarskoj zoni, iako nije klasifikovao sve (ili većinu) instanci tačno. Ovaj primer ilustruje odnos ustaljenih mera evaluacije modela i njihovu primenu u drugim oblastima, gde, iako evaluacija modela nije savršena, model pruža značajne informacije.

Implikacije za buduća istraživanja su značajne. Daljim unapređivanjem modela, verovatno će se dostići tačka gde dodatne promene modela, pretraga hiperparametara i dodavanje novih atributa neće doneti značajna poboljšanja. U tom slučaju, moguće je sprovesti dodatna istraživanja koja bi se fokusirala na lažno pozitivne instance, gde bi pretraga ofiolita koji se nalaze ispod površine bila od značaja. Sa druge strane, ovaj pristup istraživanja može se primeniti i na druge litologije, gde je ključno detektovati položaj ciljne litologije pod zemljom, kao što su određena ležišta mineralnih sirovina i drugi resursi što otvara nove mogućnosti za primenu raznovrsnih metoda kao što su mašinsko učenje.

U prikazanom primeru, pet najinformativnijih atributa uključivalo je dva atributa koja nisu uobičajeno dostupna velikoj grupi istraživača. Naime, karta Bugeovih anomalija i karta totalnog intenziteta Zemljinog magnetnog polja redukovanog na pol predstavljaju veoma značajne podatke. Dobijanje ovih podataka, ako za dato područje nisu već izmereni, može predstavljati značajnu prepreku prilikom planiranja istraživanja, kako zbog intenzivnosti u pogledu vremena potrebnog za njihovo dobijanje, potrebnih instrumenata, terenskog vremena i ekonomičnosti.

Buduća istraživanja, osim što će se fokusirati na razvijanje same metode i primenu na druge litologije, mogu imati za cilj razvoj mera i predloga za postizanje najboljih mogućih rezultata uz minimalnu potrošnju resursa. Ovakav pristup neće težiti stvaranju modela koji tačno klasifikuje što veći broj instanci, već modela koji predstavlja vrlo dobru aproksimaciju sa što manje utrošenih resursa. Ovaj pristup je značajan jer omogućava široj zajednici istraživača da primene datu metodu, čak i onima koji nemaju pristup specifičnim podacima, poput gravimetrijskih i magnetometrijskih podataka.

Takođe, istraživanje zasnovano na otvorenim podacima, kao što je slučaj sa atributom karte udaljenosti od reka, koji je dobijen sa portala otvorenih geo-prostornih podataka (OpenStreetMaps), takođe je od velikog značaja. Razvijanje metode koja će biti dostupna široj zajednici istraživača, koji nemaju pristup određenim podacima, u kombinaciji sa razvojem metoda zasnovanih na otvorenim podacima, smatra se od ključne važnosti.

4.3. Detektovanje anomalija amplitude VLF signala i prognoziranje talasovodnih parametara niske jonosfere Zemlje

Poglavlje koje prikazuje primenu modela vođenih podacima na podatke amplitude VLF signala i aproksimaciju talasovodnih parametara Zemljine niske jonosfere ima vrlo široku primenu za buduća istraživanja. Podpoglavlje koje se odnosi na detekciju anomalije amplitude VLF signala (3.3.1.) ima najširi spektar podataka koji se mogu koristiti za treniranje modela. Javno dostupna baza podataka WALDO (eng. *Worldwide Archive of Low-Frequency Data and Observations*) sadrži odabrane podatke amplitude VLF signala i faze za period od 2005. do 2017. godine.

Planirana buduća istraživanja na ovu temu su jednim delom već započeta, formiranjem i postavljanjem okvira standardizacije (eng. *Standardization framework*). Prvi problem odnosi se na nadgledano mašinsko učenje, gde je potrebno da model ima označene sve instance u trening setu podataka, što zahteva ljudske resurse za ručno označavanje podataka. Drugi problem je razvoj odgovarajućih pratećih alata za nesmetano preuzimanje, skadištenje, transformaciju i obradu podataka iz WALDO baze podataka.

Buduća istraživanja će se fokusirati na razvoj alata za nesmetano preuzimanje, skladištenje i transformaciju podataka koji će omogućiti ručno označavanje. Kada se prikupi dovoljan broj označenih podataka, moguće je preći sa binarne na višeklasnu klasifikaciju i dalji razvoj

modela. Takođe, sa obzirom na sveobuhvatnost podataka u WALDO bazi, istraživanje zasnovano na širem angažovanju zajednice (eng. *Community- based research*) može pružiti značajne koristi. Velike korporacije ulažu značajna sredstva u različite vrste mašinskog učenja, a timovi ljudi često rade isključivo na obeležavanju podataka za nadgledano mašinsko učenje. Pošto se ovakav pristup u ovom istraživanju ne može koristiti na isti način kao u velikim korporacijama, šira istraživačka zajednica može značajno ubrzati označavanje podataka. Zbog toga je uloženo dosta truda u razvijanje okvira standardizacije, sa primerima kako optimalno obeležiti podatke amplitude VLF signala. Takođe, prethodno pomenuti alati biće razvijeni kao otvorenog koda, dostupni svima.

Drugi primer prikazan u ovom poglavlju odnosi se na aproksimaciju talasovodnih parametara D-sloja jonosfere (3.3.2.). Istraživanje je pokazalo da je moguće razviti algoritam koji daje dovoljno dobru aproksimaciju talasovodnih parametara, uz to što su u ovom primeru korišćene neke nestandardne prakse u okviru mašinskog učenja. Najpre, trening set podataka je bio aproksimiran iz manjeg skupa podataka, a taj manji set je takođe korišćen za testiranje podataka. Ovaj pristup, iako nije standardan, primenjen je uz oprez kako bi rezultati ostali validni.

Primena najboljeg modela na validacionom setu podataka pokazala je da je moguće razviti adekvatnu metodu za aproksimaciju talasovodnih parametara, ali da buduća istraživanja moraju biti fokusirana na pribavljanje adekvatnog seta podataka za trening, bez potrebe za primenom metoda preuzorkovanja podataka.

4.4. Primena statističkih metoda na podatke magnetne susceptibilnosti uzoraka sa jalovišta rudnika „Rudnik“

Jedini primer u okviru disertacije koji se nije eksplicitno ticao metoda mašinskog učenja ili prognoziranja vremenskih serija bio je primer primene statističkih metoda na podatke magnetne susceptibilnosti sa uzoraka sa jalovišta rudnika „Rudnik“. Ovaj primer bio je interesantan iz više razloga. Najpre, prikazao je da su merenja magnetne susceptibilnosti u niskom polju vrlo efikasna i ekonomična za dobijanje raspodele magnetne susceptibilnosti, a samim tim i teških metala po dubini tela jalovišta. Sa druge strane, ovaj primer se uklapa u širi kontekst disertacije jer prikazuje sličnu primenu podataka- izvlačenje dodatnih informacija iz

već izmerenih podataka i testiranje alternativnih metoda koje se konvencionalno ne primenjuju u datoj oblasti.

Primena metoda vremenskih serija na prostorno zavisne podatke dala je različite rezultate. Na primer, metode za detekciju stacionarnosti u prostornim serijama potencijalno se mogu povezati sa vrstom deponovanja materijala. Ova tvrdnja zvuči veoma interesantno, ali su potrebna dodatna istraživanja sa različitim primerima koji pokrivaju kako antropogenu, tako i prirodnu depoziciju materijala kako bi se verifikovala. Druge primenjene metode, kao što je grafikon prve diference, nisu pružile dodatne informacije koje nisu već očigledne vizuelnom analizom originalnih vrednosti magnetne susceptibilnosti. Sa druge strane, grafikon kumulativne sume pokazuje potencijal za dalja istraživanja zbog svoje dobre korelacije sa prevojnim tačkama u bušotinama i granicama slojeva, ali su i za ovo potrebni dodatni primeri.

Primena drugih metoda, kao što je Kolmogorov- Smirnov test, dala je pozitivan rezultat, gde se u velikoj meri slojevi izdvojeni samo analizom vrednosti magnetne susceptibilnosti odlično poklapaju sa litološkim podacima. Nastavak istraživanja, sa podacima iz dodatne četiri bušotine sa bedema flotacijskog jalovišta, biće značajan za dalju verifikaciju izvedenih zaključaka.

Primer prikazuje vrlo interesantnu situaciju koja osvetljava teškoće primene geološke, ali i statističke logike prilikom analize podataka koji dolaze od uzoraka sa antropogeno deponovanog materijala. Kao što je prethodno bilo reči, materijal je lokalnog porekla, ali se tokom procesa prerade rude međusobno mešao do trenutka deponovanja, što čini bilo kakvu geološku, a samim tim i statističku logiku veoma izazovnom. Nastavak ovakve vrste istraživanja materijal sa jalovišta, kao i drugih materijala trebalo bi da omoguće jasnije sagledavanje navedenih izazova. Pored toga, primena statističkih metoda iz drugih oblasti ima potencijal da pruži dodatne informacije koje nisu dostupne konvencionalnim metodama.

5. Zaključak

Tema doktorske disertacije odnosi se na primenu modela vođenih podacima u geofizici, fizici atmosfere i jonosfere. U okviru teme prikazano je nekoliko primera koji se odnose na različite primene metoda mašinskog učenja, prognoziranja i analize vremenskih serija, kao i statističkih metoda, kako bi se dobile dodatne informacije iz podataka koje konvencionalne metode ne omogućavaju, kao i za automatizaciju određenih procesa koji se uglavnom vrše od strane istraživača i predstavljaju vrlo vremenski intenzivne procese.

Iako primeri obuhvataju različite naučne oblasti, glavni cilj disertacije je identifikacija područja u kojima se modeli vođeni podacima mogu uspešno primeniti. U okviru disertacije predstavljeno je šest primera iz četiri različite naučne oblasti, gde su modeli vođeni podacima kao i šira oblast nauke o podacima korišćeni za različite svrhe. Raznovrsnost naučnih disciplina, vrsta podataka, namena i ciljeva dodatno ističe primenljivost modela vođenih podacima, kao i nauke o podacima, te naglašava značaj njihovog daljeg razvoja i primene u budućnosti u oblasti geonauka.

Koncentracija zagađujućih materija u vazduhu predstavlja vrlo značajan problem za širu javnost, kao i za određene grupe stanovništva koje boluju od respiratornih i kardiovaskularnih bolesti. Industrijalizacija, ekspanzija gradova, upotreba motornih vozila, kao i sagorevanje materijala za grejanje domaćinstava, doprinose problemu povećane koncentracije zagađujućih materija u vazduhu. Prvi primer odnosio se na primenu Fejsbukovog Profet algoritma za prognozu budućih vrednosti koncentracije zagađujućih materija u vazduhu. Primer je pokazao da je Fejsbukov Profet algoritam sposoban za kvalitetne prognoze, ali da su potrebna dodatna istraživanja kako bi se dobio model koji je stabilan, pouzdan i prikazuje minimalna odstupanja prilikom prognoza. Sa druge strane, kontinuirani monitoring koncentracije zagađujućih materija u vazduhu može biti opterećen tehničkim problemima, što može dovesti do preskočenih opservacija. Drugi primer odnosio se na primenu metoda mašinskog učenja za imputaciju preskočenih podataka. Analiza dobijenih vrednosti pokazala je da je osnovna ideja dvosmernog imputovanja podataka dobra, ali da je potrebno dodatno istraživanje, jer je razvijeni algoritam pokazao vrednosti uporedive sa najjednostavnijim modelima, uz povećan utrošeni računarski resurs.

Prostorna klasifikacija ofiolita istočne Vardarske zone predstavlja primer klasifikacije geološke jedinice za koju su primenjeni kombinovani geološki, geofizički, satelitski i drugi podaci. U prvoj iteraciji istraživanja dobijeni su rezultati koji pokazuju da je model, u dobrom smislu, naučio generalni prostorni položaj ofiolita, iako nisu sve instance ofiolita bile tačno klasifikovane. Druga iteracija istraživanja uključivala je zamenu primenjene metode, unapređenje metode pretraživanja hiperparametara i dodavanje drugih atributa. U drugoj fazi istraživanja, jedna od glavnih mera evaluacije modela (F1- mera) prikazala je skoro duplo povećanje usled zamene modela i dodavanjem dodatnog atributa. Buduća istraživanja biće usmerena ka pronalaženju dodatnih atributa i daljem proširenju prostora i metoda.

Prikazana su i dva primera koji su se odnosili na analizu podataka VLF signala i propagacionih parametara niske jonosfere Zemlje. Prvi primer odnosio se na automatizaciju detekcije anomalija jonosferskog amplitude VLF signala. Istraživanje je pokazalo da su primenjene metode imale svojih prednosti i da su u određenim slučajevima adekvatno klasifikovale anomalni VLF signal. Sa druge strane, postojali su i slučajevi u kojima je primenjena metoda dala vrlo loše klasifikacije, što je i prikazano u disertaciji. Buduća istraživanja fokusiraće se na ostvarivanje standarda za multiklasnu klasifikaciju različitih karakteristika VLF signala, a potom i na prelazak sa binarne na višeklasnu klasifikaciju mašinskim učenjem. Određivanje talasovodnih parametara jonosfere predstavlja kompleksan proces koji u određenim situacijama može dati loše ocene talasovodnih parametara. Drugi primer prikazuje primenu metoda mašinskog učenja za modelovanje talasovodnih parametara tokom poremećenih jonosferskih stanja. Rezultati su pokazali da, iako je obrada i modelovanje bila ograničena kvantitetom podataka, postoji mogućnost razvijanja modela za modelovanje talasovodnih parametara primenom mašinskog učenja. Pošto je inicijalno istraživanje koristilo metode progušćivanja za dobijanje adekvatnog uzorka za dato modelovanje, buduća istraživanja treba da se fokusiraju na prikupljanje kvalitetnog seta podataka za treniranje modela sa dovoljnim brojem instanci.

Na kraju, primenjene statističke metode na podatke magnetne susceptibilnosti izmerene u niskom polju na uzorcima sa flotacijskog jalovišta prikazale su potencijalnu primenu tih metoda u datom području. Metode analize vremenskih serija primenjene na prostorno zavisne podatke imaju svoje benefite, kao što su metode ocenjivanja stacionarnosti koje se potencijalno mogu dovesti u vezu sa načinom deponovanja. Za dobijanje pouzdanih informacija od takvih metoda potrebno je više istraživanja. Sa druge strane, statističke metode su pokazale svoju

dobru stranu pre svega prilikom određivanja da li dati, izdvojeni sloj ima statistička svojstva koja su različita od prethodnih i narednih slojeva.

Disertacija je obuhvatila široku primenu datih metoda u različitim poljima, uključujući geofiziku, fiziku atmosfere i fiziku jonosfere, kao i za različite vrste podataka. Buduća istraživanja u okviru svih prikazanih primera, kao i u potrazi za novim primerima gde modeli vođeni podacima mogu doneti unapređenja, biće nastavljena daljim istraživanjima.

6. Literatura

Ahmedov, B., Mirzaev, B., Mamatov, F., Khodzhaev, D., Julliev, M., 2020. Integrating of GIS and GPS for ionospheric perturbations in D- and F-layers using VLF receiver. InterCarto InterGIS 26, 547–560. https://doi.org/10.35595/2414-9179-2020-1-26-547-560

Akhavi, M.S., Webster, T.L., Raymond, D.A., 2001. RADARSAT-1 Imagery and GIS Modeling for Mineral Exploration in Nova Scotia, Canada. Geocarto International 16, 57–64. https://doi.org/10.1080/10106040108542183

Albert, G., Ammar, S., 2021. Application of random forest classification and remotely sensed data in geological mapping on the Jebel Meloussi area (Tunisia). Arabian Journal of Geosciences 14. https://doi.org/10.1007/s12517-021-08509-x

Aliyu, A., Adamu, L.M., Abdulmalik, N.F., Amuda, A.K., Umar, A.O., Umar, N., Jungudo, S.M., 2021. Application Of Remote Sensing in Lithological Discrimination of Precambrian Basement Rocks of Zungeru Area, Part of Sheet 163 (Zungeru Nw), North Central Nigeria. FUDMA Journal of Sciences 5, 390–398. https://doi.org/10.33003/fjs-2021-0503-729

Al-Rawashdeh, S., Saleh, B., Hamzah, M., 2006. The use of Remote Sensing Technology in geological Investigation and mineral Detection in El Azraq-Jordan. Cybergeo. https://doi.org/10.4000/cybergeo.2856

Alsaber, A.R., Pan, J., Al-Hurban, A., 2021. Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018). International Journal of Environmental Research and Public Health 18, 1333. https://doi.org/10.3390/ijerph18031333

Araujo, J.A., 2010. Particulate air pollution, systemic oxidative stress, inflammation, and atherosclerosis. Air Quality Atmosphere & Health 4, 79–93. https://doi.org/10.1007/s11869-010-0101-8

Arnaut F., 2024. River Proximity Data as A Predictor for Ophiolite Classification: A Machine Learning Approach with OSM Data. XII Congress of BGS GEOPHYSICS FOR THE BETTER WORLD, Kopaonik Mt., Republic of Serbia, 27-31 May, 2024 (saopštenje štampano u celini sa kongresa- zbornik neobjavljen).

Arnaut, F., Cvetkov, V., Đurić, D., Samardžić-Petrović, M., 2023a. Short-term forecasting of PM10 and PM2.5 concentrations with Facebook's Prophet Model at the Belgrade-Zeleno brdo. Geofizika 40. https://doi.org/10.15233/gfz.2023.40.7

Arnaut, F., Đurđević, V., Kolarski, A., Srećković, V.A., Jevremović, S., 2024a. Improving Air Quality Data Reliability through Bi-Directional Univariate Imputation with the Random Forest Algorithm. Sustainability 16, 7629. https://doi.org/10.3390/su16177629

Arnaut, F., Đurić, D., Đurić, U., Samardžić-Petrović, M., Peshevski, I., 2024b. Application of geophysical and multispectral imagery data for predictive mapping of a complex geo-tectonic unit: a case study of the East Vardar Ophiolite Zone, North-Macedonia. Earth Science Informatics 17, 1625–1644. https://doi.org/10.1007/s12145-024-01243-4

Arnaut, F., Kolarski, A., Srećković, V.A., 2023b. Random Forest Classification and Ionospheric Response to Solar Flares: Analysis and Validation. Universe 9, 436. https://doi.org/10.3390/universe9100436

Arnaut, F., Kolarski, A., Srećković, V.A., 2024c. Machine Learning Classification Workflow and Datasets for Ionospheric VLF Data Exclusion. Data 9, 17. https://doi.org/10.3390/data9010017

Arnaut, F., Kolarski, A., Srećković, V.A., Mijić, Z., 2023c. Ionospheric Response on Solar Flares through Machine Learning Modeling. Universe 9, 474. https://doi.org/10.3390/universe9110474

Arnaut, F., Kolarski, A., Srećković, V.A., Langović, M., Jevremović, S. 2025. Standardization Framework of Ionospheric Very Low Frequency (VLF) Signal Amplitude Classes for Machine Learning-Based Anomaly Detection: From Calm Ionospheric Conditions to Solar Activity-Induced Dynamics. Contributions of the Astronomical Observatory Skalnaté Pleso (u procesu publikacije)

Batista, G.E. a. P.A., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6, 20–29. https://doi.org/10.1145/1007730.1007735

Behnia, P., Harris, J.R., Rainbird, R.H., Williamson, M.C., Sheshpari, M., 2012. Remote predictive mapping of bedrock geology using image classification of Landsat and SPOT data, western Minto Inlier, Victoria Island, Northwest Territories, Canada. International Journal of Remote Sensing 33, 6876–6903. https://doi.org/10.1080/01431161.2012.693219

Belaschsen, I., Broday, D.M., 2022. Imputation of Missing PM2.5 Observations in a Network of Air Quality Monitoring Stations by a New kNN Method. Atmosphere 13, 1934. https://doi.org/10.3390/atmos13111934

Berger, V.W., Zhou, Y., 2014. Kolmogorov–Smirnov Test: Overview. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06558

Bernstein, J.A., Alexis, N., Barnes, C., Bernstein, I.L., Nel, A., Peden, D., Diaz-Sanchez, D., Tarlo, S.M., Williams, P.B., Bernstein, J.A., 2004. Health effects of air pollution. Journal of Allergy and Clinical Immunology 114, 1116–1123. https://doi.org/10.1016/j.jaci.2004.08.030

Bisgaard, S., Kulahci, M., 2011. Time Series Analysis and Forecasting by Example, Wiley series in probability and statistics. https://doi.org/10.1002/9781118056943

Bityukova, L., Scholger, R., Birke, M., 1999. Magnetic susceptibility as indicator of environmental pollution of soils in Tallinn. Physics and Chemistry of the Earth Part a Solid Earth and Geodesy 24, 829–835. https://doi.org/10.1016/s1464-1895(99)00122-2

Boev, B., Cvetković, V., Prelević, D., Šarić, K., Boev, I., 2018. East Vardar Ophiolites Revisited: A Brief Synthesis of Geology and Geochemical Data. Contributions Section of Natural Mathematical and Biotechnical Sciences 39, 51. https://doi.org/10.20903/csnmbs.masa.2018.39.1.119

Bolt, G.H., Bruggenwert, M.G.M., 1976. Chapter 1 Composition of the Soil, in: Developments in Psychiatry. pp. 1–12. https://doi.org/10.1016/s0166-2481(08)70630-5

Boyko, T., Scholger, R., Stanjek, H., 2004. Topsoil magnetic susceptibility mapping as a tool for pollution monitoring: repeatability of in situ measurements. Journal of Applied Geophysics 55, 249–259. https://doi.org/10.1016/j.jappgeo.2004.01.002

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/a:1010933404324

Brempong, F., Mariam, Q., Preko, K., 2016. The use of magnetic susceptibility measurements to determine pollution of agricultural soils in road proximity. African Journal of Environmental Science and Technology 10, 263–271. https://doi.org/10.5897/ajest2015.2058

Brink, H., Richards, J., Fetherolf, M., 2016. Real-World Machine Learning.

Cacciuttolo, C., Cano, D., Custodio, M., 2023. Socio-Environmental Risks Linked with Mine Tailings Chemical Composition: Promoting Responsible and Safe Mine Tailings Management Considering Copper and Gold Mining Experiences from Chile and Peru. Toxics 11, 462. https://doi.org/10.3390/toxics11050462

Carranza, E.J.M., Laborte, A.G., 2014. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Computers & Geosciences 74, 60–70. https://doi.org/10.1016/j.cageo.2014.10.004

Chen, M., Zhu, H., Chen, Y., Wang, Y., 2022. A Novel Missing Data Imputation Approach for Time Series Air Quality Data Based on Logistic Regression. Atmosphere 13, 1044. https://doi.org/10.3390/atmos13071044

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minin 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21. https://doi.org/10.1186/s12864-019-6413-7

Chougule, A., Chamola, V., Sam, A., Yu, F.R., Sikdar, B., 2023. A Comprehensive Review on Limitations of Autonomous Driving and Its Impact on Accidents and Collisions. IEEE Open Journal of Vehicular Technology 5, 142–161. https://doi.org/10.1109/ojvt.2023.3335180

Cracknell, M.J., Reading, A.M., 2013. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Computers & Geosciences 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Cvetkov V., Đurić D., Lesić V., Starčević M., Petković M,. Petrović S., 2016. Koenigsberger ratio and Total Magnetic Field Anomaly reduction to the pole for the area of Macedonia. Geologica Macedonica 4:429–534

Çorbacıoğlu, Ş.K., Aksel, G., 2023. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turkish Journal of Emergency Medicine 23, 195–198. https://doi.org/10.4103/tjem.tjem_182_23

Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O., Ajibuwa, O.E., 2019. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5, e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

Dimitrijević MD (1978) Geološko kartiranje. Izdavački i informativni studentski centar, Beograd.

Dockery, D.W., Schwartz, J., Spengler, J.D., 1992. Air pollution and daily mortality: Associations with particulates and acid aerosols. Environmental Research 59, 362–373. https://doi.org/10.1016/s0013-9351(05)80042-8

Dumurdžanov N, Hristov S, Pavlovski B, Ivanova V (1981) Tumač za listove Vitolište i Kajmakčalan. Osnovna geološka karta (1:100.000) Socijalističke Federativne Republike Jugoslavije. Savezni geološki zavod, Beograd, str. 61 (na makedonskom).

Farrand, W.H., 1997. Identification and mapping of ferric oxide and oxyhydroxide minerals in imaging spectrometer data of Summitville, Colorado, U.S.A., and the surrounding San Juan Mountains. International Journal of Remote Sensing 18, 1543–1552. https://doi.org/10.1080/014311697218269

Fedrizzi, M., de Paula, E.R., Kantor, I.J., Langley, R.B., Santos, M.C., 2002. Mapping the low-latitude ionosphere with GPS. GPS WORLD 2002, 13, 41–47.

Ferguson, J. Computer Programs for Assessment of Long-Wavelength Radio Communications, Version 2.0: User's Guide and Source Files; Space and Naval Warfare Systems Center: San Diego, CA, USA, 1998.

Flores, A., Tito-Chura, H., Centty-Villafuerte, D., Ecos-Espino, A., 2023. Pm2.5 Time Series Imputation with Deep Learning and Interpolation. Computers 12, 165. https://doi.org/10.3390/computers12080165

Foody, G.M., Mathur, A., 2004. A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on Geoscience and Remote Sensing 42, 1335–1343. https://doi.org/10.1109/tgrs.2004.827257

Ge, Y.-Z., Zhang, Z.-J., Cheng, Q.-M., Wu, G.-P., 2021. Geological mapping of basalt using stream sediment geochemical data: Case study of covered areas in Jining, Inner Mongolia, China. Journal of Geochemical Exploration 232, 106888. https://doi.org/10.1016/j.gexplo.2021.106888

Gómez-García, C., Martin-Hernandez, F., García, J.Á.L., Martínez-Pagán, P., Manteca, J.I., Carmona, C., 2015. Rock magnetic characterization of the mine tailings in Portman Bay (Murcia, Spain) and its contribution to the understanding of the bay infilling process. Journal of Applied Geophysics 120, 48–59. https://doi.org/10.1016/j.jappgeo.2015.06.008

Ham, J., Chen, N.Y., Crawford, M.M., Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 43, 492–501. https://doi.org/10.1109/tgrs.2004.842481

Hanesch, M., Scholger, R., 2005. The influence of soil type on the magnetic susceptibility measured throughout soil profiles. Geophysical Journal International 161, 50–56. https://doi.org/10.1111/j.1365-246x.2005.02577.x

Harris, J.R., Ford, K.L., Charbonneau, B.W., 2009. Application of gamma-ray spectrometer data for lithological mapping in a cordilleran environment, Sekwi Region, NWT. Canadian Journal of Remote Sensing 35, S12–S30. https://doi.org/10.5589/m09-022

Harris, J.R., Grunsky, E.C., 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Computers & Geosciences 80, 9–25. https://doi.org/10.1016/j.cageo.2015.03.013

Harris, J.R., He, J.X., Rainbird, R., Behnia, P., 2014. A Comparison of Different Remotely Sensed Data for Classifying Bedrock Types in Canada's Arctic: Application of the Robust Classification Method and Random Forests. Geoscience Canada 41, 557. https://doi.org/10.12789/geocanj.2014.41.062

Harris, J.R., Rogge, D., Hitchcock, R., Ijewliw, O., Wright, D., 2005. Mapping lithology in Canada's Arctic: application of hyperspectral data using the minimum noise fraction transformation and matched filtering. Canadian Journal of Earth Sciences 42, 2173–2193. https://doi.org/10.1139/e05-064

Harris, J.R., Schetselaar, E.M., De Kemp, E., St-Onge, M.R., 2008. Case study 2. LANDSAT, magnetic and topographic data for regional lithological mapping, southeast Baffin Island. https://doi.org/10.4095/226015

Hasanin, T., Khoshgoftaar, T., 2018. The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. 2018 IEEE International Conference on Information Reuse and Integration (IRI) 70–79. https://doi.org/10.1109/iri.2018.00018

Hristov S, Karajovanović M, Stračkov M (1965) Osnovna geološka karta SFRJ, list Kavadarci, M 1:100.000 (karta i tumač). Savezni geološki zavod, Beograd, str. 62 (na makedonskom).

Hristov S, Karajovanović M, Stračkov M (1973) Osnovna geološka karta bivše Jugoslavije 1:100.000, tumač za list Kavadarci (na makedonskom).

Huang, C., Davis, L.S., Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing 23, 725–749. https://doi.org/10.1080/01431160110040323

Ivanovski T, Rakićević T (1966) Osnovna geološka karta bivše Jugoslavije 1:100.000, list Gevgelija. Savezni geološki zavod, Beograd (na makedonskom).

Jaffar, S.T.A., Chen, L.-Z., Younas, H., Ahmad, N., 2017. Heavy metals pollution assessment in correlation with magnetic susceptibility in topsoils of Shanghai. Environmental Earth Sciences 76. https://doi.org/10.1007/s12665-017-6598-5

Jiang, N., Li, Y., Zuo, H., Zheng, H., Zheng, Q., 2020. BiLSTM-A: A missing value imputation method for PM2.5 prediction. 2020 2nd International Conference on Applied Machine Learning (ICAML) 23–28. https://doi.org/10.1109/icaml51583.2020.00014

Jordanova, D., Goddu, S.R., Kotsev, T., Jordanova, N., 2012. Industrial contamination of alluvial soils near Fe–Pb mining site revealed by magnetic and geochemical studies. Geoderma 192, 237–248. https://doi.org/10.1016/j.geoderma.2012.07.004

Joshi, M.V., 2003. On evaluating performance of classifiers for rare classes. Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference On 641–644. https://doi.org/10.1109/icdm.2002.1184018

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Karajovanović M, Hristov S (1976) Tumač Osnovne geološke karte Kumanova 1:100.000, list karte: Skoplje. Savezni geološki zavod Jugoslavije, str. 58 (na makedonskom, sa rezimeom na engleskom).

Karajovanović, M, Hadži-Mitrova S (1975) Osnovna geološka karta bivše Jugoslavije 1:100.000, tumač za list Titov Veles. Savezni geološki zavod, Beograd (na makedonskom).

Karamata, S., 2006. The geological development of the Balkan Peninsula related to the approach, collision and compression of Gondwanan and Eurasian units. Geological Society London Special Publications 260, 155–178. https://doi.org/10.1144/gsl.sp.2006.260.01.07

Karimi, R., Ayoubi, S., Jalalian, A., Sheikh-Hosseini, A.R., Afyuni, M., 2011. Relationships between magnetic susceptibility and heavy metals in urban topsoils in the arid region of Isfahan, central Iran. Journal of Applied Geophysics 74, 1–7. https://doi.org/10.1016/j.jappgeo.2011.02.009

Kebalepile, M.M., Dzikiti, L.N., Voyi, K., 2024. Using Diverse Data Sources to Impute Missing Air Quality Data Collected in a Resource-Limited Setting. Atmosphere 15, 303. https://doi.org/10.3390/atmos15030303

Kim, J., Jung, S.P., Chul, M.C., Youn, S.L., 2010. Relationship between magnetic susceptibility and heavy metal content of soil. In *19th World Congress of Soil Science, Soil Solutions for a Changing World*. pp. 1-6.

Kim, T., Kim, J., Yang, W., Lee, H., Choo, J., 2021. Missing Value Imputation of Time-Series Air-Quality Data via Deep Neural Networks. International Journal of Environmental Research and Public Health 18, 12213. https://doi.org/10.3390/ijerph182212213

Kuhn, S., Cracknell, M.J., Reading, A.M., 2018. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. Geophysics 83, B183–B193. https://doi.org/10.1190/geo2017-0590.1

Kumar, S., Kumar, A., Menk, F., Maurya, A.K., Singh, R., Veenadhari, B., 2014. Response of the low-latitude D region ionosphere to extreme space weather event of 14–16 December 2006. Journal of Geophysical Research Space Physics 120, 788–799. https://doi.org/10.1002/2014ja020751

Lecoanet, H., Lévêque, F., Segura, S., 1999. Magnetic susceptibility in environmental applications: comparison of field probes. Physics of the Earth and Planetary Interiors 115, 191–204. https://doi.org/10.1016/s0031-9201(99)00066-7

Leverington, D.W., 2010. Discrimination of sedimentary lithologies using Hyperion and Landsat Thematic Mapper data: a case study at Melville Island, Canadian High Arctic. International Journal of Remote Sensing 31, 233–260. https://doi.org/10.1080/01431160902882637

Leverington, D.W., Moon, W.M., 2012. Landsat-TM-Based Discrimination of Lithological Units Associated with the Purtuniq Ophiolite, Quebec, Canada. Remote Sensing 4, 1208–1231. https://doi.org/10.3390/rs4051208

Libasin, Z., Ul-Saufie, A.Z., Ahmat, H., Shaziayani, W.N., 2020. Single and Multiple Imputation Method to Replace Missing Values in Air Pollution Datasets: A Review. IOP Conference Series Earth and Environmental Science 616, 012002. https://doi.org/10.1088/1755-1315/616/1/012002

Liu, D., Liu, Z., Wang, Y., Zhou, L., 2023. Editorial: Understanding heavy metal pollution and control in the environment around metal tailings. Frontiers in Environmental Science 11. https://doi.org/10.3389/fenvs.2023.1168949

Longhi, I., Sgavetti, M., Chiari, R., Mazzoli, C., 2001. Spectral analysis and classification of metamorphic rocks from laboratory reflectance spectra in the 0.4-2.5 μ m interval: A tool for hyperspectral data interpretation. International Journal of Remote Sensing 22, 3763–3782. https://doi.org/10.1080/01431160010006980

Lorenz, H., 2004. Integration of Corona and Landsat Thematic Mapper data for bedrock geological studies in the high Arctic. International Journal of Remote Sensing 25, 5143–5162. https://doi.org/10.1080/01431160410001705097

Hossin, M., Sulaiman, M.N., 2015. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process 5, 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Meyler, A., Kenny, G. and Quinn, T., 1998. Forecasting Irish inflation using ARIMA models, Technical paper 3. *Economic Analysis, Research and Publications Department, Central Bank of Ireland, PO Box, 559*.

Mitra, A.P., 1978. The D-region of the ionosphere. Endeavour 2, 12–21. https://doi.org/10.1016/0160-9327(78)90028-5

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., Mian, A., 2023. A Comprehensive Overview of Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.06435

Neville, R.A., Lévesque, J., Staenz, K., Nadeau, C., Hauff, P., Borstad, G.A., 2003. Spectral unmixing of hyperspectral imagery for mineral exploration: comparison of results from SFSI and AVIRIS. Canadian Journal of Remote Sensing 29, 99–110. https://doi.org/10.5589/m02-085

Nikolić, M., Žečević, A., 2019. Mašinsko učenje. Univerzitet u Beogradu, Matematički fakultet. (Skripta za predmet Mašinsko učenje).

Nišić, D., Aleksić, N., Živanović, B., Pantelić, U., Rupar, V., 2024. Review of the Failure at the Flotation Tailings Storage Facility of the “Stolice” Mine (Serbia). Applied Sciences 14, 10163. https://doi.org/10.3390/app142210163

Norazian, M.N., Shukri, Y.A., Azam, R.N., Bakri, A.M.M.A., 2008. Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 34, 341. https://doi.org/10.2306/scienceasia1513-1874.2008.34.341

Ohya, H., Nishino, M., Murayama, Y., Igarashi, K., Saito, A., 2005. Using tweek atmospherics to measure the response of the low-middle latitude D-region ionosphere to a magnetic storm. Journal of Atmospheric and Solar-Terrestrial Physics 68, 697–709. https://doi.org/10.1016/j.jastp.2005.10.014

Oudeika, M.S., Altinoglu, F.F., Akbay, F., Aydin, A., 2020. The use of magnetic susceptibility and chemical analysis data for characterizing heavy metal contamination of topsoil in Denizli city, Turkey. Journal of Applied Geophysics 183, 104208. https://doi.org/10.1016/j.jappgeo.2020.104208

Papastefanopoulos, V., Linardatos, P., Kotsiantis, S., 2020. COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. Applied Sciences 10, 3880. https://doi.org/10.3390/app10113880

Pendžerkovski J, Rakićević T, Ivanovski T, Gjuzelkovski D (1963) Geološka karta i tumač za list Kožuf (K 34–105), Osnovna geološka karta SFRJ 1:100.000. Savezni geološki zavod (na makedonskom, sa rezimeom na engleskom). Beograd, str. 47.

Petrović D., (2015) Prostorni položaj ofiolita istočne Vardarske zone: geofizičko-geološki model i njegove geodinamičke implikacije. Univerzitet u Beogradu (na srpskom, sa rezimeom na engleskom), Rudarsko-geološki fakultet.

Petrovský, E., Kapička, A., Jordanova, N., Knab, M., Hoffmann, V., 2000. Low-field magnetic susceptibility: a proxy method of estimating increased pollution of different environmental systems. Environmental Geology 39, 312–318. https://doi.org/10.1007/s002540050010

Rahman, E.Ab., Hamzah, F.M., Latif, M.T., Azid, A., 2023. Forecasting PM2.5 in Malaysia Using a Hybrid Model. Aerosol and Air Quality Research 23, 230006. https://doi.org/10.4209/aaqr.230006

Rakholia, R., Le, Q., Vu, K., Ho, B.Q., Carbajo, R.S., 2022. AI-based air quality PM2.5 forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. Urban Climate 46, 101315. https://doi.org/10.1016/j.uclim.2022.101315

Rakićević T, Dumurdžanov N, Petkovski P (1969) Geološka karta i tumač za list Štip (K 34–81), Osnovna geološka karta SFRJ 1:100.000. Savezni geološki zavod (na makedonskom, sa rezimeom na engleskom). Beograd, str. 70.

Rakićević T, Pendžerkovski J, Kovačevik M (1973) Geološka karta i tumač za list Strumica (K 34–94), Osnovna geološka karta SFRJ 1:100.000. Savezni geološki zavod (na makedonskom, sa rezimeom na engleskom). Beograd, str. 69.

Rakićević T, Stojanov R, Arsovski M (1965) Geološka karta i tumač za list Prilep (K 34–92), Osnovna geološka karta SFRJ 1:100.000. Savezni geološki zavod (na makedonskom, sa rezimeom na engleskom). Beograd, str. 65.

Robertson, A., Karamata, S., Šarić, K., 2008. Overview of ophiolites and related units in the Late Palaeozoic–Early Cenozoic magmatic and tectonic development of Tethys in the northern part of the Balkan region. Lithos 108, 1–36. https://doi.org/10.1016/j.lithos.2008.09.007

Salehi, M.H., Jorkesh, Sh., Mohajer, R., 2013. Relationship between Magnetic Susceptibility and Heavy Metals Concentration in Polluted Soils of Lenjanat Region, Isfahan. E3S Web of Conferences 1, 04003. https://doi.org/10.1051/e3sconf/20130104003

Samal, K.K.R., Babu, K.S., Das, S.K., Acharaya, A., 2019. Time Series based Air Pollution Forecasting using SARIMA and Prophet Model. Proceedings of the 2019 International Conference on Information Technology and Computer Communications 80–85. https://doi.org/10.1145/3355402.3355417

Samardžić- Petrović, M., (2014) Predicting land use change with data-driven models. Univerzitet u Beogradu (na engleskom sa rezimeom na srpskom), Građevinski fakultet.

Saripuddin, M., Suliman, A., Sameon, S.S., Jorgensen, B.N., 2021. Random Undersampling on Imbalance Time Series Data for Anomaly Detection. Proceedings of the 2021 the 4th International Conference on Machine Learning and Machine Intelligence 151–156. https://doi.org/10.1145/3490725.3490748

Schetselaar, E.M., Ryan, J., 2008. Case study 9. A remote predictive mapping case study of the Boothia mainland area, Nunavut, Canada. https://doi.org/10.4095/226028

Schmid, S.M., Bernoulli, D., Fügenschuh, B., Matenco, L., Schefer, S., Schuster, R., Tischler, M., Ustaszewski, K., 2008. The Alpine-Carpathian-Dinaridic orogenic system: correlation and evolution of tectonic units. Swiss Journal of Geosciences 101, 139–183. https://doi.org/10.1007/s00015-008-1247-3

Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126, 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Shakeel, A., Chong, D., Wang, J., 2023a. Load forecasting of district heating system based on improved FB-Prophet model. Energy 278, 127637. https://doi.org/10.1016/j.energy.2023.127637

Shakeel, A., Chong, D., Wang, J., 2023b. District heating load forecasting with a hybrid model based on LightGBM and FB-prophet. Journal of Cleaner Production 409, 137130. https://doi.org/10.1016/j.jclepro.2023.137130

Shen, J., Valagolam, D., McCalla, S., 2020. Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea. PeerJ 8, e9961. https://doi.org/10.7717/peerj.9961

Simić, V., Petrović, S., Arnaut, F., Cvetkov, V., Kostović, M., Radulović, D., Stojanović, J., Jovanović, V., Todorović, D., Nikolić, N., Senćanski, J., Bogdanović, G. & Marilović, D. 2024. 'Projekat PRIZMA: Karakterizacija i tehnološki postupci za reciklažu i

ponovnu upotrebu flotacijske jalovine rudnika „Rudnik"', *Zapisnici Srpskog geološkog društva*. Beograd: Srpsko geološko društvo.

Somyanonthanakul, R., Warin, K., Amasiri, W., Mairiang, K., Mingmalairak, C., Panichkitkosolkul, W., Silanun, K., Theeramunkong, T., Nitikraipot, S., Suebnukarn, S., 2022. Forecasting COVID-19 cases using time series modeling and association rule mining. BMC Medical Research Methodology 22. https://doi.org/10.1186/s12874-022-01755-x

Sposito G., 1989. The Chemistry of Soils. Oxford University Press

Srećković, V.A., Šulić, D.M., Ignjatović, L., Vujčić, V., 2021a. Low Ionosphere under Influence of Strong Solar Radiation: Diagnostics and Modeling. Applied Sciences 11, 7194. https://doi.org/10.3390/app11167194

Srećković, V.A., Šulić, D.M., Vujčić, V., Mijić, Z.R., Ignjatović, L.M., 2021b. Novel Modelling Approach for Obtaining the Parameters of Low Ionosphere under Extreme Radiation in X-Spectral Range. Applied Sciences 11, 11574. https://doi.org/10.3390/app112311574

Su, C., Rana, N.M., Zhang, S., Wang, B., 2024. Environmental pollution and human health risk due to tailings storage facilities in China. The Science of the Total Environment 928, 172437. https://doi.org/10.1016/j.scitotenv.2024.172437

Tejasvini, K.N., Amith, G.R., Akhtharunnisa, N., Shilpa, H., 2020. Air Pollution Forecasting Using Multiple Time Series Approach, in: Advances in Intelligent Systems and Computing. pp. 91–100. https://doi.org/10.1007/978-981-15-2188-1_8

Vasiliev, A., Gorokhova, S., Razinsky, M., 2020. Technogenic Magnetic Particles in Soils and Ecological–Geochemical Assessment of the Soil Cover of an Industrial City in the Ural, Russia. Geosciences 10, 443. https://doi.org/10.3390/geosciences10110443

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., Van Mulbregt, P., Vijaykumar, A., Pietro Bardelli, A., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.-L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P.,

Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., De Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wait, J.R.; Spies, K.P., 1964. Characteristics of the Earth-Ionosphere Waveguide for VLF Radio Waves; US Department of Commerce, National Bureau of Standards: Gaithersburg MD, USA, Volume 13

Wang, X.S., 2013. Assessment of heavy metal pollution in Xuzhou urban topsoils by magnetic susceptibility measurements. Journal of Applied Geophysics 92, 76–83. https://doi.org/10.1016/j.jappgeo.2013.02.015

Wardana, I.N.K., Gardner, J.W., Fahmy, S.A., 2022. Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder. Neural Computing and Applications 34, 16129–16154. https://doi.org/10.1007/s00521-022-07224-2

Waske, B., Braun, M., 2009. Classifier ensembles for land cover mapping using multitemporal SAR imagery. ISPRS Journal of Photogrammetry and Remote Sensing 64, 450–457. https://doi.org/10.1016/j.isprsjprs.2009.01.003

Wijesekara, W.M.L.K.N., Liyanage, L., 2020. Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index, in: Advances in Intelligent Systems and Computing. pp. 257–269. https://doi.org/10.1007/978-3-030-39442-4_20

Ye, Z., 2019. Air Pollutants Prediction in Shenzhen Based on ARIMA and Prophet Method. E3S Web of Conferences 136, 05001. https://doi.org/10.1051/e3sconf/201913605001

ZainEldin, H., Gamel, S.A., El-Kenawy, E.-S.M., Alharbi, A.H., Khafaga, D.S., Ibrahim, A., Talaat, F.M., 2022. Brain Tumor Detection and Classification Using Deep Learning and Sine-Cosine Fitness Grey Wolf Optimization. Bioengineering 10, 18. https://doi.org/10.3390/bioengineering10010018

Zawadzki, J., Fabijańczyk, P., Magiera, T., Rachwał, M., 2015. Geostatistical Microscale Study of Magnetic Susceptibility in Soil Profile and Magnetic Indicators of Potential Soil Pollution. Water Air & Soil Pollution 226. https://doi.org/10.1007/s11270-015-2395-5

Zhang, B., Shi, H., Wang, H., 2023. Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. Journal of Multidisciplinary Healthcare Volume 16, 1779–1791. https://doi.org/10.2147/jmdh.s410301

Zhang, Y., Sun, Q., Liu, J., Petrosian, O., 2023. Long-Term Forecasting of Air Pollution Particulate Matter (PM2.5) and Analysis of Influencing Factors. Sustainability 16, 19. https://doi.org/10.3390/su16010019

Zhang, Z., Zhang, S., 2023. Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks. International Journal of Environmental Science and Technology 20, 13535–13550. https://doi.org/10.1007/s13762-023-04900-1

Zhoul, L., Chenl, M., Ni, Q., 2020. A hybrid Prophet-LSTM Model for Prediction of Air Quality Index. 2021 IEEE Symposium Series on Computational Intelligence (SSCI) 595–601. https://doi.org/10.1109/ssci47803.2020.9308543

Žunić, E., Korjenić, K., Hodžić, K., Đonko, D., 2020. Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-world Data. International Journal of Computer Science and Information Technology 12, 23–36. https://doi.org/10.5121/ijcsit.2020.12203

Zuo, R., Carranza, E.J.M., 2023. Machine Learning-Based Mapping for Mineral Exploration. Mathematical Geosciences 55, 891–895. https://doi.org/10.1007/s11004-023-10097-3

Biografija

Filip Arnaut rođen je 14. jula 1997. godine u Beogradu, gde je završio osnovnu i srednju elektrotehničku školu „Rade Končar“, smer elektrotehničar elektronike. Osnovne akademske studije na Univerzitetu u Beogradu, na Rudarsko-geološkom fakultetu, studijski program Geofizika, upisuje 2016. godine, a završava 2020. godine sa prosečnom ocenom 8,29/10,00 i odbranjenim diplomskim radom na temu „*Primena ukrštenog kvadratnog dispozitiva za detekciju primarne zone rasta korena drveta vrbe*“, ocenjenim ocenom 10,00.

U septembru 2020. godine upisao je master akademske studije na istom fakultetu (studijski program Geofizika), koje završava u julu 2021. godine sa prosečnom ocenom 9,36/10,00 i odbranjenim master radom na temu „*Korelabilnost solarnog vetra sa seizmičkim događajima u zoni Balkanskog poluostrva*“, takođe ocenjenim ocenom 10,00. U oktobru 2021. godine upisuje doktorske akademske studije na Rudarsko-geološkom fakultetu Univerziteta u Beogradu, na studijskom programu Geologija, sa usmerenjem na oblasti Geofizike, gde je ispite predviđene doktorskim studijama položio sa prosečnom ocenom 10,00/10,00.

Istraživačko zvanje istraživač-pripravnik stiče u septembru 2022. godine na Departmanu za geofiziku, Rudarsko-geološkog fakulteta, Univerziteta u Beogradu. Od maja 2023. godine zaposlen je u Institutu za fiziku u Beogradu, Univerziteta u Beogradu, kao istraživač-pripravnik u Laboratoriji za astrofiziku i fiziku jonosfere. Autor je ili koautor ukupno 10 publikacija sa SCI liste, a ukupno 45 radova objavljenih u domaćim i međunarodnim časopisima ili izloženih na domaćim i međunarodnim konferencijama. Takođe je obavljao dužnosti urednika zbornika radova sa međunarodnih konferencija, kao i dužnosti predsednika ili sekretara organizacionog odbora na međunarodnim skupovima. Dobitnik je više nagrada i stipendija za naučno-istraživački rad.

Izjave

Изјава о ауторству

Име и презиме аутора Филип Арнаут

Број индекса Г804/21

Изјављујем

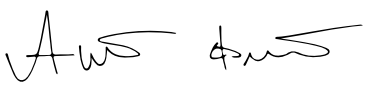
да је докторска дисертација под насловом

Примена модела вођених подацима у геофизици

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, _____



Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Филип Арнаут

Број индекса Г804/21

Студијски програм Геологија

Наслов рада Примена модела вођених подацима у геофизици

Ментор др Весна Цветков, редовни професор

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, _____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Примена модела вођених подацима у геофизици

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

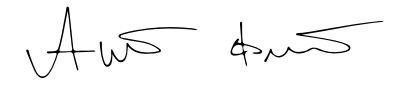
5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци. Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, _____



1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.