Razvoj prognostičkog modela GPS ionosferskog kašnjenja s komponentama geomagnetskog polja kao prediktorima, metodama strojnog učenja

Petranović, Mihael

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski sveučilišni studij računarstva

Diplomski rad

Razvoj prognostičkog modela GPS ionosferskog kašnjenja s komponentama geomagnetskog polja kao prediktorima, metodama strojnog učenja

Rijeka, rujan 2024.

Mihael Petranović
0069085309

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski sveučilišni studij računarstva

Diplomski rad

Razvoj prognostičkog modela GPS ionosferskog kašnjenja s komponentama geomagnetskog polja kao prediktorima, metodama strojnog učenja

Mentor: prof. dr. sc. Renato Filjar

Rijeka, rujan 2024.

Mihael Petranović
0069085309

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA DIPLOMSKE ISPITE

Rijeka, 13. ožujka 2023.

Zavod: **Zavod za računarstvo**
Predmet: **Programski određen radio**
Grana: **2.09.03 obradba informacija**

ZADATAK ZA DIPLOMSKI RAD

Pristupnik: **Mihael Petranović (0069085309)**
Studij: Sveučilišni diplomski studij računarstva
Modul: Programsko inženjerstvo

Zadatak: **Razvoj prognostičkog modela GPS ionosferskog kašnjenja s komponentama geomagnetskog polja kao prediktorima, metodama strojnog učenja / Development of a machine learning-based predictive model of GPS ionospheric delay with geomagnetic field components as p**

Opis zadatka:

Ionosfersko kašnjenje signala satelitskih navigacijskih sustava predstavlja najznačajniji pojedinačni uzrok pogreške satelitskog određivanja položaja. Štetni učinci ionosferskog kašnjenja se prevladavaju različitim postupcima, uključujući modele ispravaka (korekcijske modele). U okviru ovog rada je potrebno razviti prognostički model ionosferskog kašnjenja za Global Positioning System (GPS) zasnovan na zadanim eksperimentalnim opažanjima ionosferskog kašnjenja i opažanjima vrijednosti komponenata lokalnog geomagnetskog polja prikupljanih cjelodnevno tokom godine dana u sub-ekvatorijalnom području, uz primjenu metoda strojnog učenja. Potrebno je prikazati problem nastanka i učinaka ionosferskog kašnjenja te teorijske postavke korištenih metoda strojnog učenja, napraviti opisnu statističku analizu podataka (komponenata geomagnetskog polja kao prediktora i ionosferskog kašnjenja kao rezultirajućeg ishoda), različitim metodama strojnog učenja razviti barem tri kandidata modela, te postupkom provjere uspješnosti odrediti optimalni model korekcije GPS ionosferskog kašnjenja. Komentirati načine primjenjivosti razvijenog modela. Istraživanje je potrebno provesti u programskom okruženju za statističko računarstvo R.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 20. ožujka 2023.

Mentor:

Prof. dr. sc. Renato Filjar

Predsjednik povjerenstva za diplomski ispit:

Prof. dr. sc. Miroslav Joler

Izjava o samostalnoj izradi rada

Izjavljujem da sam samostalno izradio ovaj rad.

Rijeka, rujan 2024.

_____
Ime Prezime

Zahvala

Zahvaljujem mentoru prof dr. sc. Renatu Filjaru na korisnim raspravama, savjetima i vodstvu prilikom izrade diplomskog rada. Zahvaljujem obitelji na podršci tijekom studiranja.

Sadržaj

Popis slika

Popis tablica

Poglavlje 1

Uvod

Satelitska navigacija jedna je od ključnih komponenti modernog društva zbog svoje upotrebe u raznim područjima, od transporta i logistike do vojnih operacija i znanstvenih istraživanja [1]. Glavnu ulogu ovdje imaju globalni navigacijski satelitski sustavi (eng. Global Navigation Satellite System (GNSS)), kojih je u trenutku pisanja rada četvero;

- Američki Globalni Položajni Sustav
  (eng. Global Positioning System (GPS)) [2]
- Europski Galileo [3]
- Ruski GLONASS
  (rus. GLObalnaya NAvigazionnaya Sputnikovaya Sistema) [4]
- Kineski BeiDou [5]

Svaki navedeni GNSS sastoji se od satelitske konstelacije, mreže satelita u orbiti koje pružaju usluge globalnog određivanja položaja, navigacije i mjerenja vremena (eng. Positioning, Navigation and Timing (PNT)), koristeći se signalima iz svemira preko kojih odašilju podatke o vremenu i rasponu odgovarajućim prijemnicima, gdje se ti

podatci koriste za određivanje lokacije [6].

Točnost i preciznost dobivenih mjerenja nije u potpunosti ispravna jer ovisi o mnogim čimbenicima poput pozicije satelita, greške u unutarnjim satovima satelita, šumu u signalu te atmosferskim i geomagnetskim uvjetima [7]. Od svih navedenih, najveći utjecaj na satelitsko određivanje položaja ima ionosfersko kašnjenje zbog kojeg se, u slučajima rada s jednofrekventnim prijemnicima, pri završnom određivanju položaja dobivena vrijednost može razlikovati i za nekoliko desetina metara od stvarne [8]. Ionosfersko kašnjenje fenomen je koji nastaje prolaskom signala kroz ionosferu, jedan od slojeva Zemljine atmosfere bogat električki nabijenim česticama, što uzrokuje promjene u brzini i putanji signala te naposljetku i grešku u određivanju položaja. Ovisi o geografskoj lokaciji prijemnika, intenzitetu solarne aktivnosti i dobu dana, a mjeri se kao ukupni sadržaj elektrona (eng. Total Electron Content (TEC)). Razvijanje standardnog globalnog korektivnog modela za ionosfersko kašnjenje (poput NeQuicka kojeg koristi Galileo [9] ili Klobucharovog modela kod GPS-a [10]) ili regionalnih korektivnih modela te razvoj naprednih satelitskih metoda pozicioniranja, kao što su kinematika u stvarnom vremenu (eng. Real-time Kinematics (RTK)) i preciznog pozicioniranja točki (eng. Precise Point Positioning (PPP)), neki su od pristupa kojima se pokušava ublažiti učinak ionosferskog kašnjenja [8].

U sklopu rada opisano je, razvijeno i objašnjeno nekoliko prognostičkih modela GPS ionosferskog kašnjenja pomoću metoda strojnog učenja.

Diplomski rad sastoji se od sedam poglavlja. Uvodno poglavlje općenito definira tematiku rada i ukazuje na negativne posljedice ionosferskog kašnjenja. U drugom poglavlju opisana su prethodna istraživanja na danu temu koja se detaljnije dotiču ionosferskog kašnjenja, Klobucharovog modela i strojnog učenja. Treće poglavlje predstavlja tezu rada; moguća je izrada ciljanog modela ionosferskih korekcija korištenjem postupaka strojnog učenja. Kod četvrtog poglavlja cjelovito je opisana metodologija rada kao i način vrednovanja izrađenih modela. Peto poglavlje prikazuje

rezultate rada izrađenih modela, dok su u šestom poglavlju ti rezultati analizirani i protumačeni. Posljednje, sedmo poglavlje, sažima dosad odrađeni rad i spominje moguća buduća proširenja.

Poglavlje 2

Prethodna istraživanja

2.1 Ionosfersko kašnjenje

Ionosferom smatramo dio gornje Zemljine atmosfere koji se proteže na udaljenosti od približno 50 kilometara do više od 1000 kilometara, gdje je primjetan utjecaj na širenje radio valova zbog mogućnosti prisustva dovoljne ionizacije čestica [11]. Prolaskom GNSS satelitskog signala kroz ionosferu dolazi do kašnjenja signala što rezultira pogreškama u određivanju položaja, a sam iznos kašnjenja ovisi o ukupnom sadržaju elektrona, izraženog u TEC jedinicama, TECU (eng. Total Electron Content Units), gdje je jedan TECU jednak 10^{16} elektrona po kvadratnom metru (matematički e^{-}/m^{2}) [8]. Ionosfersko kašnjenje nije konstantno i varira ovisno o nekolicini čimbenika, uključujući:

- Doba dana - jača solarna radijacija tijekom dana uzrokuje povećanje ionizacije u atmosferi čime se povećava TEC, dok noću ionizacija opada zbog slabije solarne radijacije
- Sezonske promjene - tijekom ljeta ionosferska aktivnost je jača zbog veće solarne radijacije pa su i razine TEC-a veće

- Solarni ciklus - posebice solarni maksimum, razdoblje najveće solarne aktivnosti koje se redovito javlja tijekom Sunčevog jedanaestogodišnjeg solarnog ciklusa, kada su povećanja TEC-a najizraženija
- Geografska širina - ekvatorijalna područja podložnija su većim razinama TEC-a zbog jače solarne radijacije
- Geomagnetske oluje - u polarnim područjima, smještenim u visokim geografskim širinama, razine TEC-a mogu značajno varirati zbog geomagnetskih oluja uzrokovanih interakcijama solarnih vjetrova sa Zemljinim magnetskim poljem

Za jednofrekventni GNSS prijamnik, TEC uzrokuje kašnjenje signala, tj. pogrešku u mjerenju ΔR na frekvenciji nosivog vala f koja se može prikazati matematičkim izrazom (2.1) [8], [12].

$$\Delta R = 40,3 \cdot \frac{TEC}{f^2} \tag{2.1}$$

2.2 Klobucharov model

Klobucharov model je standardni korektivni model za ionosfersko kašnjenje izrađen za američki GNSS, GPS, temeljen na empirijskom pristupu istoimenog autora u članku objavljenom 1987. godine [13]. Osmišljeni algoritam za rad koristi osam koeficijenata prenesenih putem navigacijskih poruka koje korisnicima odašilju sateliti GPS-a.

Glavna prednost modela je niska računalna zahtjevnost i jednostavnost, što je ključno kako bi se mogao primjenjivati u realnom vremenu zbog ograničene količine računalnih resursa unutar GNSS prijemnika. Navedena jednostavnost dolazi pod cijenu nepouzdanosti modela u slučajevima gdje je visoka preciznost potrebna, odnosno kada su varijacije TEC-a značajne; tijekom solarnih maksimuma i u uvjetima visoke geomagnetske aktivnosti. Tada dolazi do grešaka u procjeni ionosferskog

kašnjenja koje mogu rezultirati višemetarskim greškama određivanja položaja, što je veliki nedostatak u današnje vrijeme gdje je visokoprecizno pozicioniranje od iznimne važnosti.

Klobucharov model smanjuje korijen srednje kvadratne pogreške (eng. Root Mean Square Error (RMSE)) predviđanja ionosferskog kašnjenja za približno 55% [13] i bez obzira na to što postoje moderniji i točniji modeli, kao što je NeQuick koji se koristi u Galileu ili napredniji dvofrekventni pristup koji je moguć na dvofrekventnim GNSS prijemnicima, Klobucharov model se i danas upotrebljava zbog svoje jednostavnosti i nasljedne integracije (eng. Legacy Integration) s modernim bazama podataka i aplikacijama.

2.3 Strojno učenje

Strojno učenje (eng. Machine Learning (ML)) grana je računalne znanosti i umjetne inteligencije (eng. Artificial Intelligence (AI)) koja se usredotočuje na razvijanje umjetne inteligencije u smjeru oponašanja ljudskog učenja, korištenjem potrebnih podataka i algoritama kako bi postepeno poboljšavala točnost učenja [14]. Algoritmi strojnog učenja treniraju modele korištenjem velikih količina podataka, s ciljem prepoznavanja odnosa i obrazaca unutar istih te njihovu primjenu na nove, nepoznate slučajeve, zbog stečene sposobnosti donošenja odluka u smislu predviđanja kategoričkih klasa (klasifikacija) ili kontinuiranih brojčanih vrijednosti (regresija). Spomenute algoritme može se podijeliti u nekoliko kategorija [15]:

- Nadzirano učenje (eng. Supervised Learning) - model je treniran na skupu podataka koji je podijeljen u ulazne i izlazne podatke, a zadaća mu je naučiti model ispravnom predviđanju izlaza za nove, nepoznate podatke
- Nenadzirano učenje (eng. Unsupervised Learning) - model je treniran na skupu podataka koji nisu eksplicitno određeni ili klasificirani pa je zadaća modela

otkrivanje skrivenih obrazaca i struktura unutar danih podataka

- Polu-nadzirano učenje (eng. Semi-supervised learning) - model koristi klasificirane i neklasificirane podatke kako bi na temelju njih generirao funkciju predviđanja ili klasifikacije
- Pojačano učenje (eng. Reinforcement Learning) - model donosi odluke za koje dobiva povratne informacije na temelju kojih se postepeno optimizira

Postoje mnogobrojni algoritmi, kao i njihove modificirane i kombinirane verzije, koji se koriste u strojnom učenju, a neki od njih su linearna regresija (eng. Linear Regression), stroj potpornih vektora (eng. Support Vector Machine (SVM)), stablo odluke (eng. Decision Tree), slučajna šuma (eng. Random Forest), K-najbliži susjed (eng. K-Nearest Neighbors), gradijentno pojačavanje (eng. Gradient Boosting), neuronske mreže (eng. Neural Network) itd. Svaki od navedenih algoritama nudi različite pristupe za rješavanje problema, a odabir koji algoritam koristiti donosi se na temelju prirode podataka i zadanom zadatku.

Zbog mogućnosti modeliranja složenih i nelinearnih odnosa između različitih faktora koji utječu na ukupan sadržaj elektrona u ionosferi, metode strojnog učenja i njihove primjene u praksi tema su mnogih znanstvenih radova. Unutar znanstvenog članka [16] opisana je izrada nekoliko varijanti modela gradijentnog pojačavanja koji predviđaju vrijednosti vertikalnog TEC-a, ukupnog sadržaja elektrona iznad određene točke na Zemljinoj površini, uzimajući varijable koje predstavljaju solarnu aktivnost, solarni vjetar, geomagnetsku aktivnost, međuplanetarno magnetno polje, godišnje doba i doba dana kao prediktore. Autori rada [17] su statističkim metoda učenja razvili tri modela (model linearne regresije, višeslojne neuronske mreže i slučajne šume) za predviđanje vrijednosti TEC-a namijenjenih za rad u umjerenim geomagnetskim uvjetima, koristeći opažanja komponenata geomagnetskog polja kao prediktorima, a naglasili su kako je visoka raznolikost uspješnosti izvedbe modela

vjerojatno uzrokovana oskudnošću korištenih podataka. U članku [18] predstavlja se prilagođeni model, tzv. *Prophet* koji spaja indirektne metode prognoziranja s metodama strojnog učenja te koristi koeficijente sferne harmonijske funkcije u svrhu stvaranja globalne karte predviđenih vrijednosti TEC-a. Sukladno navedenim izvorima, priloženi modeli mogu osigurati veću robusnost signala GNSS-a od standardnih modela korekcije u uvjetima visokih geomagnetskih aktivnosti i solarnih radijacija te smanjiti pogreške u pozicioniranju.

Poglavlje 3

Teza rada

Za poboljšanje performansi GNSS-a, kako je prethodno istaknuto, ključna je precizna predikcija ionosferskog kašnjenja. Nakon razmatranja ograničenja tradicionalnih korekcijskih modela ionosferskog kašnjenja poput Klobucharevog, čija točnost opada tijekom sezonskih promjena, regionalnih anomalija i u uvjetima visokih geomagnetskih aktivnosti, prelazak na razvoj modela uz pomoć strojnog učenja podrazumijeva se iz nekoliko razloga:

- Sposobnost obrade velikih količina podataka - algoritmi strojnog učenja sposobni su analizirati i učiti iz ogromnih skupova podataka koji uključuju povijesne varijable TEC-a, solarne i geomagnetske aktivnosti i dr., bez ograničenja što se tiče maksimalnog broja varijabli potrebnih za izračun, rezultirajući preciznijim predviđanjima ionosferskog kašnjenja
- Bolja generalizacija - naprednije generaliziranje obrazaca u podatcima i pronalaženje skrivenih veza među varijablama omogućava općenitiju primjenu modela i njegovu bolju funkcionalnost u različitim geografskim regijama
- Prilagodljivost - nestaje potreba za korištenjem fiksnih koeficijenata jer se modeli strojnog učenja mogu kontinuirano prilagođavati promjenama uvjeta u

ionosferi, dajući mogućnost ažuriranja u stvarnom vremenu

Teza diplomskog rada temelji se na ideji da je moguće razviti prognostički model ionosferskog kašnjenja tehnikama strojnog učenja, koristeći velik broj cjelogodišnjih lokalnih opažanja komponenti gustoće geomagnetskog polja za treniranje i kasnije testiranje izrađenog modela. Model bi u teoriji mogao bit vezan uz sam GNSS prijemnik te bi direktno koristio lokalna opažanja iz okoliša u kojem se obavlja određivanja položaja, posebice komponente stanja Zemljinog magnetskog polja, čime bi potreba za vanjskim izvorima korekcija bila smanjena, odnosno složenost sustava bila bi manja. Navedene komponente geomagnetskog polja odabrane su kao prediktori jer se njihovo mjerenje jednostavno vrši preko ugrađenih osjetila u uređajima poput spomenutog prijemnika ili pametnim telefonima, odnosno nije potrebna dodatna oprema povrh već prisutne. Razvojem opisanog ionosferskog modela dobilo bi se poboljšanje točnosti određivanja položaja, a samim time direktno utjecalo i na kvalitetu svih usluga zasnovanih na satelitskoj navigaciji.

Poglavlje 4

Metodologija

4.1 Programska podrška

Prognostički modeli ionosferskog kašnjenja razvijani su korištenjem programa i alata otvorenog koda (eng. Open-source) RStudija i R jezika.

RStudio je integrirano razvojno okruženje (eng. Integrated Development Environment (IDE)) za R programski jezik koje pruža jednostavan pristup alatima za kodiranje, provjeru programskog koda, izvođenje analiza i vizualizaciju podataka [19]. Verzija korištena u radu je *2024.04.2 Build 764, Chocolate Cosmos*.

R je jezik i okruženje za statističke proračune i grafičko predstavljanje podataka [20]. Omogućuje rad s velikim skupovima podataka i korištenja tzv. R paketa, proširenja koja sadrže kod, dokumentaciju i podatke u formatu zbirke. Paketi olakšavaju i smanjuju vrijeme potrebno za pisanje programskog koda jer sadrže funkcije za statističku analizu koje se često koriste. Korištena verzija u radu je *R 4.4.0 Puppy Cup*.

4.2 Podatci

4.2.1 Dobivanje podataka

Potrebni podatci dobiveni su s repozitorija Figshare [21], čije su prikupljanje, uspoređivanje i strukturiranje u obliku strojno čitljivog CSV formata datoteke odradili autori komplementarnog znanstvenog članka [22] u kojem je detaljnije opisan cijeli proces. Dobivena datoteka sastoji se od dva skupa podataka; procijenjenih vrijednosti ukupnog sadržaja elektrona i opservacija komponenti gustoće geomagnetskog polja, svakodnevno prikupljenih tijekom 2014. godine.

Vrijednosti TEC-a procjenjuju se iz neobrađenih GPS opažanja pseudoudaljenosti, prikupljenih iz podatkovnog internetskog repozitorija internacionalnog GNSS servisa (eng. International GNSS Service (IGS)), kojeg pruža NASA [23], a dodatno su obrađeni koristeći se GPS-TEC programom [24]. Podatci su dobiveni s referentne IGS stanice u Darwinu (Sjeverni teritorij, Australija) uzorkovanjem svakih 30 sekundi.

Komponente gustoće geomagnetskog polja Bx, By i Bz, čije se vrijednosti mjere u nanoteslama (nT), preuzete su s internetskog repozitorija međunarodne mreže magnetskih opservatorija u stvarnom vremenu (eng. International Real-time Magnetic Observatory Network (INTERMAGNET)) [25]. Komponente su dobivene s referentne INTERMAGNET stanice u Kakaduu (Sjeverni teritorij, Australija) periodom uzorkovanja od jedne minute.

4.2.2 Priprema podataka

Izgled podataka unutar datoteke prikazan je na slici 4.1, sastoje se od 522 298 pojedinačnih instanci opažanja učinjenih svake minute i redom s lijeva na desno označuju:

- DOY - dan u godini (2014.)
- hr - sat u određenom danu
- min - minute u određenom satu
- sec - sekunde u određenoj minuti
- TEC - ukupni sadržaj elektrona
- dTEC - diferencijalni ukupni sadržaj elektrona
- Bx - x-komponenta gustoće geomagnetskog polja
- By - y-komponenta gustoće geomagnetskog polja
- Bz - z-komponenta gustoće geomagnetskog polja

```
DOY,hr,min,sec,TEC,dTEC,Bx,By,Bz
1,0,8,0,29.62,1.62,35434.5,2002.3,-29622.1
1,0,9,0,29.71,1.62,35434.8,2002.3,-29622
1,0,10,0,29.81,1.48,35435,2002.2,-29621.9
1,0,11,0,29.91,1.7,35435.1,2002.3,-29621.7
1,0,12,0,30.01,1.72,35435.2,2002.5,-29621.5
1,0,13,0,30.11,1.53,35435.3,2002.6,-29621.3
1,0,14,0,30.22,1.59,35435.7,2002.4,-29621.1
1,0,15,0,30.33,1.67,35436,2002.3,-29621.1
```

Slika 4.1 Isječak korištenih podataka

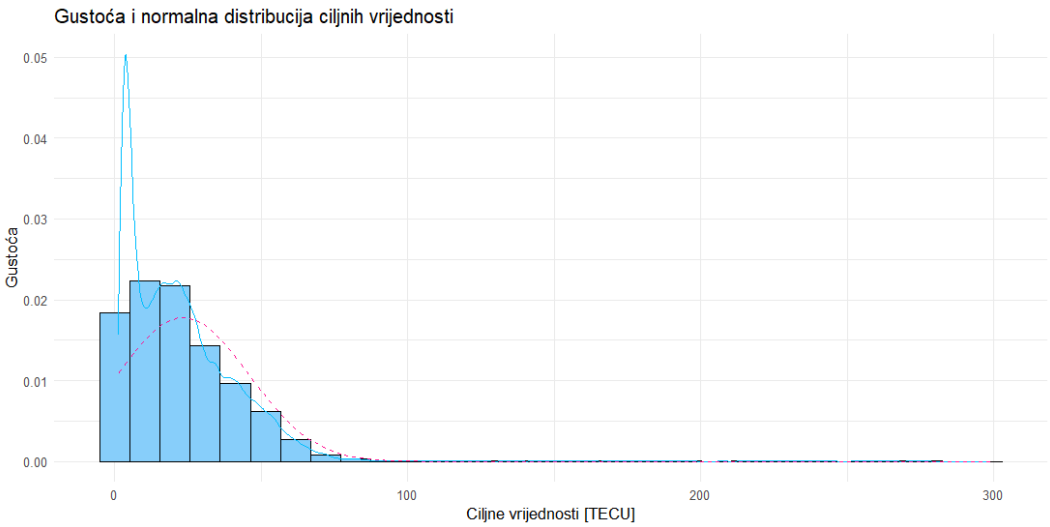
Za predviđanje tražene TEC vrijednosti kao prediktori koriste se komponente gustoće geomagnetskog polja, Bx, By i Bz, koje predstavljaju tri ortogonalna smjera geomagnetskog polja u Zemljinoj ionosferi i načelno opisuju promjene u geomagnetskom polju, dok su ostali podatci uklonjeni zbog nedovoljnog utjecaja na predviđanje kod izrade prognostičkog modela ionosferskog kašnjenja u ovom slučaju. Uklanjanjem

netipičnih vrijednosti (eng. Outliers), pojedinačnih instanci u kojima je iznos TEC-a veći od 300 (zbog uobičajenih vrijednosti u prirodi), smanjujemo skup podataka na 509 881. Opisano uređivanje podataka nužno je nakon usporedbe nekih osnovnih mjerodavnih statističkih vrijednosti, poput aritmetičke sredine, standardne varijacije i varijance, prikazanih u tablici 4.1. Zbog vidljive smanjene varijabilnost, povećane konzistencije podataka i smanjenje podatkovne raspršenosti, za razvoj prognostičkog modela uzeti su uređeni podatci kako bi procjene bile preciznije.

Tablica 4.1 Usporedba mjerodavnih vrijednosti orginalnih i uređenih podataka

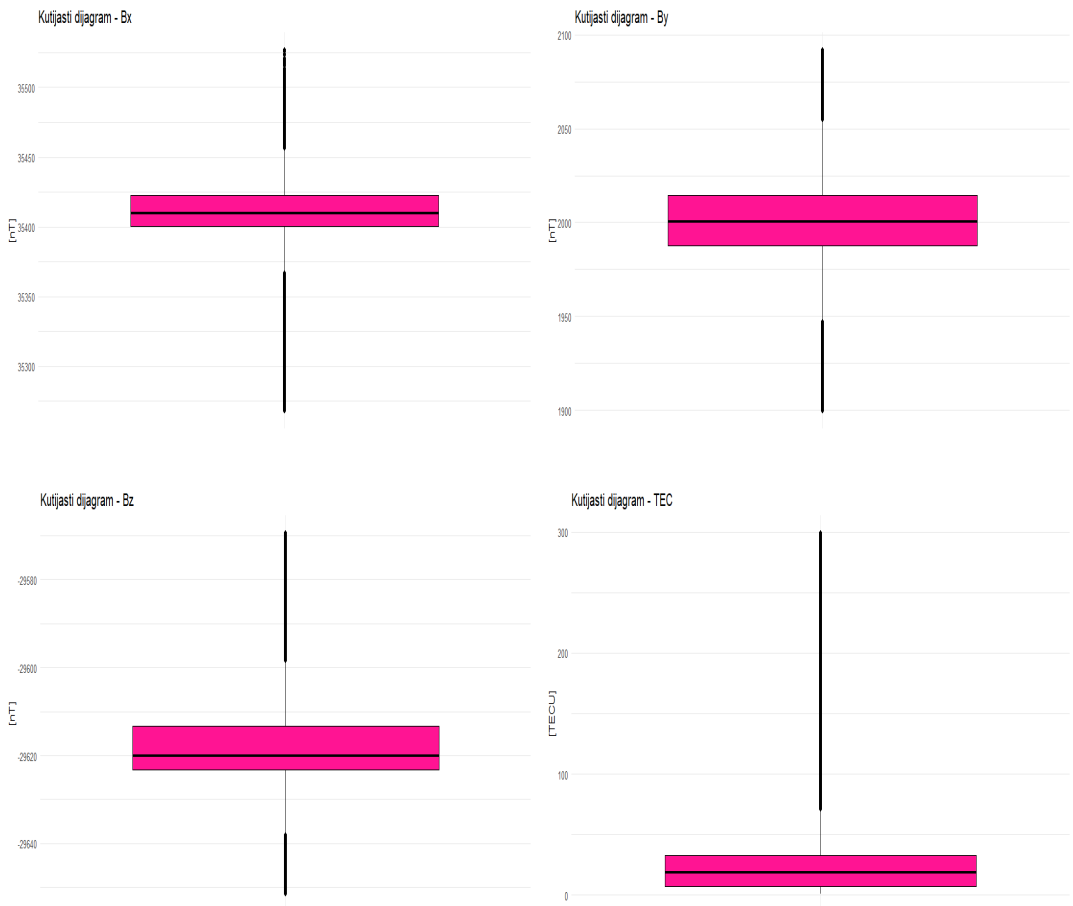
Mjerilo	Inicijalni podatci	Uređeni podatci
Aritmetička sredina	40.00449	23.17844
Std. pogreška aritm. sred.	0.162233	0.031352
Standardna devijacija	117.2459	22.38752
Std. pogreška std. dev.	0.114716	0.022169
Varijanca	13746.63	501.2021

Dijagram gustoće i normalne distribucije izmjerenih vrijednosti TEC-a na slici 4.2 daje uvid u raspodjelu vrijednosti TEC-a; asimetrična distribucija u kojoj dominiraju niže vrijednosti i koja se razlikuje od normalne distribucije (na slici prikazana ružičastom bojom), što upućuje na potrebu izrade modela koji može bolje obraditi nelinearne odnose unutar podataka i njihovu heterogenost. Na slici 4.3 može se iščitati medijan, interkvartilni raspon i netipične vrijednosti svih varijabla potrebnih za izradu modela: Bx, By, Bz i TEC. Većina podataka koncentrirana je u nižem rasponu vrijednosti (od 0 do 50 TECU), što označava kako se većina mjerenja dogodila za vrijeme stabilnih i mirnih ionosferskih uvjeta, dok su netipične vrijednosti posljedica većih geomagnetskih poremećaja.

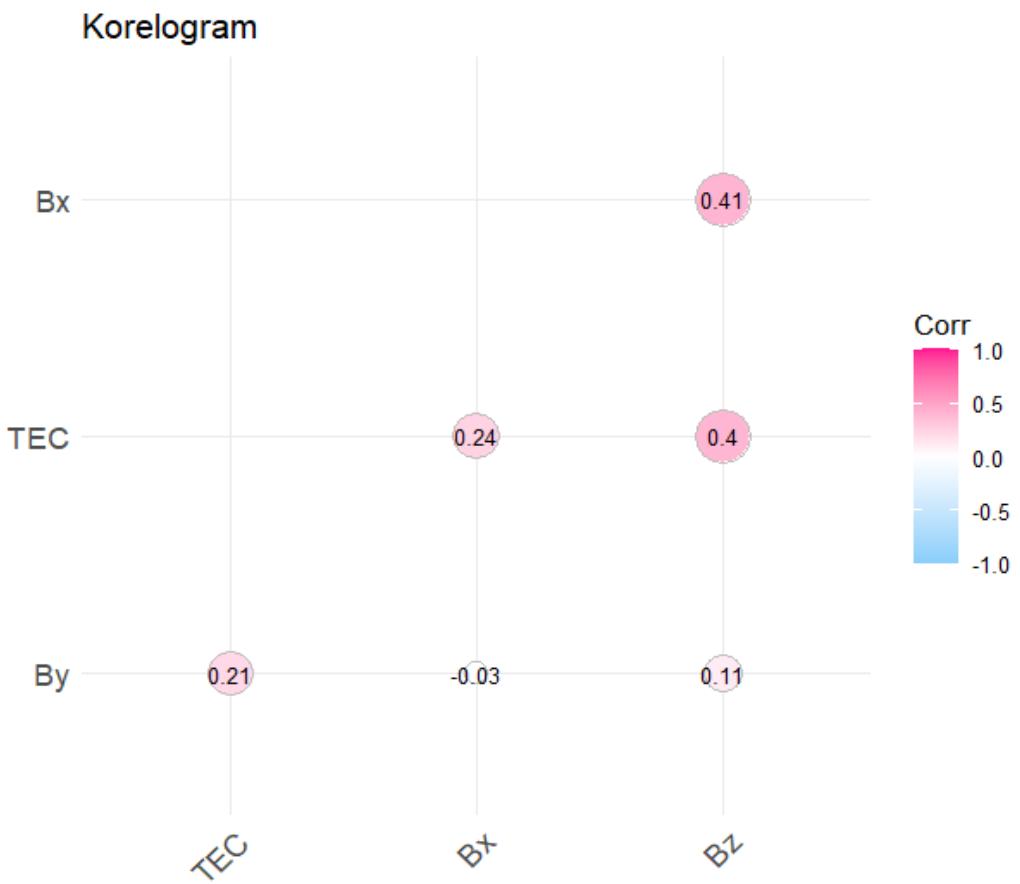


Slika 4.2 Gustoća i normalna distribucija ciljnih vrijednosti podataka

Provjerena je i međusobna korelacija varijabli čiji su rezultati vidljivi na slici 4.4. Dvije se korelacije najviše ističu, 0,41 između Bx i Bz te 0,4 između Bz i TEC, što ukazuje na umjerenu korelaciju koja je u pravilu preslaba da bi imala ozbiljniji utjecaj na razvoj i rezultate modela.



Slika 4.3 Kutijasti dijagrami podataka



Slika 4.4 Korelogram podataka

4.3 Modeli

Podatci pripremljeni na prethodno spomenuti način nasumično su podijeljeni u dva skupa; skup za treniranje i skup za testiranje, redom u omjeru 80% : 20%. Podijeljeni su kako bi se izbjegao problem pretjerane prilagodbe (eng. Overfitting), situacije u kojoj model strojnog učenja odlično predviđa vrijednosti na temelju podataka za treniranje no ne uspijeva dobro predvidjeti vrijednosti na drugačijim, novim podatcima za testiranje. Opisanim načinom podjele također se osigurava da je svaki od izrađenih modela uvježban te potom i provjeren na istom skupu podataka kako bi vrednovanje modela bilo pravedno i jednako za sve modele, bez mogućih pristranosti zbog različitih ulaznih podataka.

Razvoj prognostičkog modela m ionosferskog kašnjenja (TEC) pomoću komponenta magnetskog polja kao prediktora, Bx, By i Bz, može se generalizirati jednadžbom 4.1.

$$TEC = m(Bx, By, Bz) \tag{4.1}$$

4.3.1 Linearna regresija

Linearna regresija osnovni je statistički model strojnog učenja korišten za predviđanje kontinuiranih vrijednosti. Temelji se na modeliranju linearnog odnosa između jedne ili više nezavisnih varijabli (prediktorima) i zavisne varijable koja je ciljna, tražena vrijednost. Model linearnom regresijom pokušava pronaći koeficijente $\beta_1, \beta_2, \beta_3, ..., \beta_n$ za prediktore, minimizirajući pritom razliku između stvarnih vrijednosti i predviđanja. U slučaju većeg broja prediktora, linearni model može se prikazati jednadžbom 4.2:

$$\hat{y} = \beta_0 + \beta_1 \cdot y_1 + \beta_2 \cdot y_2 +, ..., +\beta_n \cdot y_n + \epsilon \tag{4.2}$$

gdje je:

- $\hat{y}$ - predviđena vrijednost
- β_0 - presjek s y-osi (intercept)
- $\beta_1, \beta_2, \beta_n$ - koeficijenti pridruženi stvarnim vrijednostima y (prediktorima)
- ϵ - slučajna pogreška (rezidual)

Za primjenu modela linearne regresije koristi se *lm*() funkcija iz R paketa *stats* (korištena verzija 4.4.0), jednog od osnovnih R paketa kojeg nije potrebno izričito dodat unutar programskog koda, koja se temelji na principima linearnih statističkih modela iz knjige [26]. Isprva je korišteni matematički izraz unutar funkcije izgledao kako je prikazano u jednadžbi 4.3, no detaljnijim testiranjem i usporedbom rezultata modela utvrđeno je kako malo bolje rezultate daje matematički izraz prikazan u jednadžbi 4.4. U njemu se za razliku od prethodnog dodaje međusobni interakcijski efekt između Bx i Bz kako bi model imao bolju priliku uhvatiti suptilne nelinearne odnose između te dvije komponente, koji su pretpostavljeni zbog rezultata iz prethodno odrađene korelacije podataka, a primjena navedenog u R-u prikazana je na isječku koda 4.1.

$$\text{TEC} = \beta_0 + \beta_1 \cdot Bx + \beta_2 \cdot By + \beta_3 \cdot Bz + \epsilon \tag{4.3}$$

$$\text{TEC} = \beta_0 + \beta_1 \cdot Bx + \beta_2 \cdot By + \beta_3 \cdot Bz + \beta_4 \cdot (Bx \cdot Bz) + \epsilon \tag{4.4}$$

```
linear_model <- lm(
    TEC ~ Bx + By + Bz + I(Bx * Bz),
    data = training_data
)
```

Isječak koda 4.1 Funkcija linearne regresije u R-u

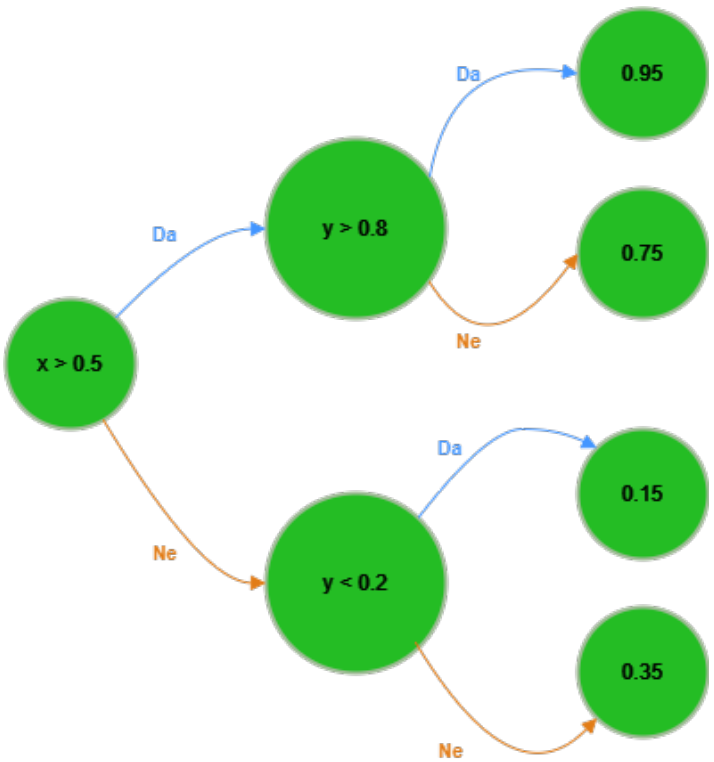
4.3.2 Stablo odluke

Stablo odluke model je koji se temelji na hijerarhijskoj strukturi koja prikazuje niz binarnih odluka, gdje svaka odluka usmjerava podatke prema jednom od dva podstabla. Koristi se pristup razlaganja podataka na sve manje podskupove putem rekurzivno primijenjenih jednostavnih odluka o podacima, koje predstavljaju čvorovi stabla, dok se ne dostigne neko konačno predviđanje koje predstavljaju listovi stabla [27]. Pojednostavljeni model ovog načina rada prikazan je na slici 4.5, gdje stablo odluke ovisno o varijablama x i y donosi jedno predviđanje.

Model stabla odluke korišten je uz pomoć R paketa (korištena verzija 4.1.23) [28] i funkcije *rpart*() unutar njega. U funkciju su poslani prediktori Bx, By i Bz te ciljna vrijednost TEC, zajedno s podatcima na kojima se model trenira, kako je prikazano na isječku koda 4.2.

```
tree_model <- rpart(
    TEC ~ Bx + By + Bz,
    data = training_data
)
```

Isječak koda 4.2 Funkcija stabla odluke u R-u

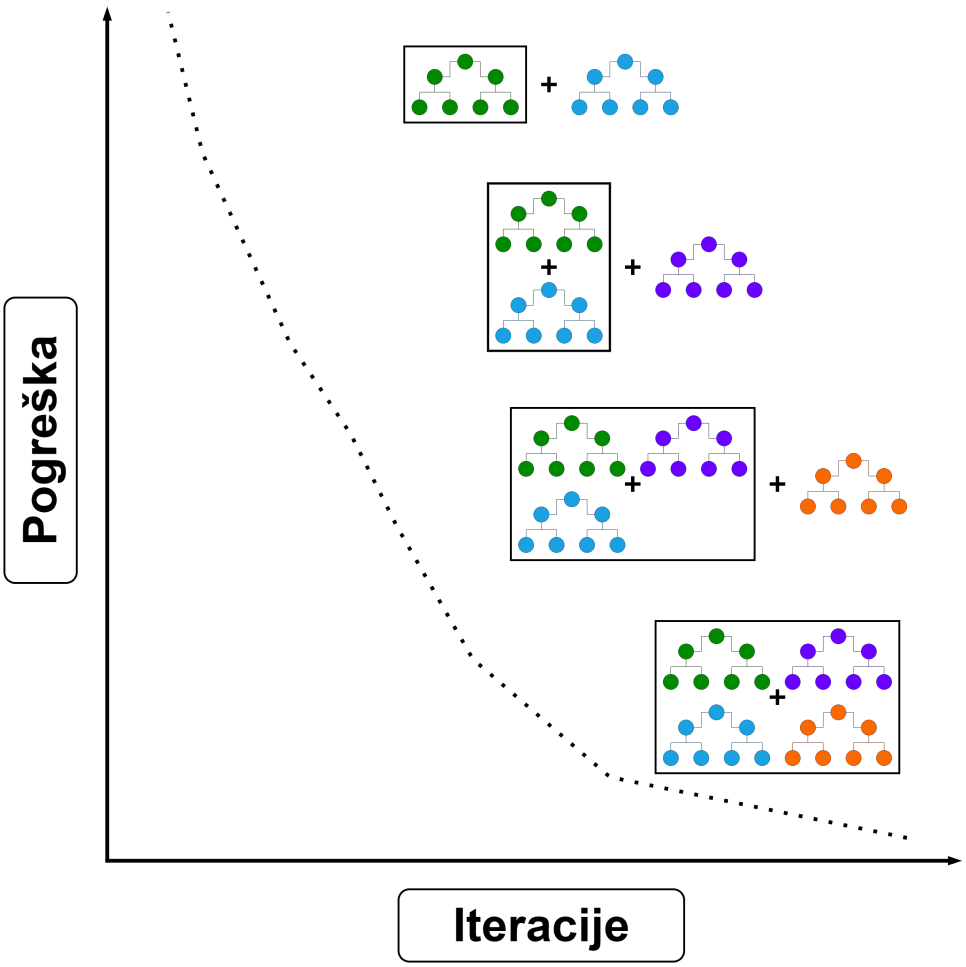


Slika 4.5 Pojednostavljeni prikaz dijagrama stabla odluke

4.3.3 Gradijentno pojačavanje

Gradijentno pojačavanje metoda je strojnog učenja koja kombinira rezultate većeg broja slabijih modela (obično jednostavnih stabala odluke) za izradu prognostičkog modela te je prvi put opisana u radu [29] gdje se još naziva i strojem za povećanje gradijenata (eng. Gradient Boosting Machine (GBM)). Koristi iterativni proces u kojem svaki od sljedećih modela u nizu pokušava ispraviti pogreške prethodnih modela, usredotočujući se na opažanja gdje su predviđanja bila najslabija, a pojednostavljeni prikaz opisanog postupka vidljiv je na slici 4.6.

Uzimajući u obzir vrijednost brzine učenja (eng. Shrinkage), kumulativnim zbrajanjem predviđanja svih stabala u modelu dolazimo do konačnih predviđanja, što



Slika 4.6 Pojednostavljeni prikaz načina rada gradijentnog pojačavanja

opisuje jednadžba 4.5:

$$\hat{y} = \sum_{m=1}^{M} \lambda \cdot f_m(x) \tag{4.5}$$

gdje je:

- $\hat{y}$ - predviđena vrijednost
- M - ukupan broj stabala

- λ - brzina učenja (shrinkage)
- f_m - predviđanja svakog stabla pomoću ulaznih prediktora x

Pristup modelu gradijentnog pojačavanja omogućuje R paket *gbm* (korištena verzija 2.2.2) preko istoimene funkcije [30], čiji je primjer korištenja prikazan na isječku koda 4.3. Parametri poslani u funkciju su redom jednadžba gdje je definirana ciljna varijabli i njeni prediktori, podatci koji će biti korišteni za treniranje modela, distribucija koja je u slučaju regresije normalna (Gaussova), broj stabala unutar modela, maksimalnu dubinu svakog stabla, brzina učenja koja kontrolira doprinos svakog stabla i naposljetku broj presavijanja kod korištenja unakrsnih provjera valjanosti (eng. Crossvalidation) za sprječavanje pretjerane prilagodbe modela.

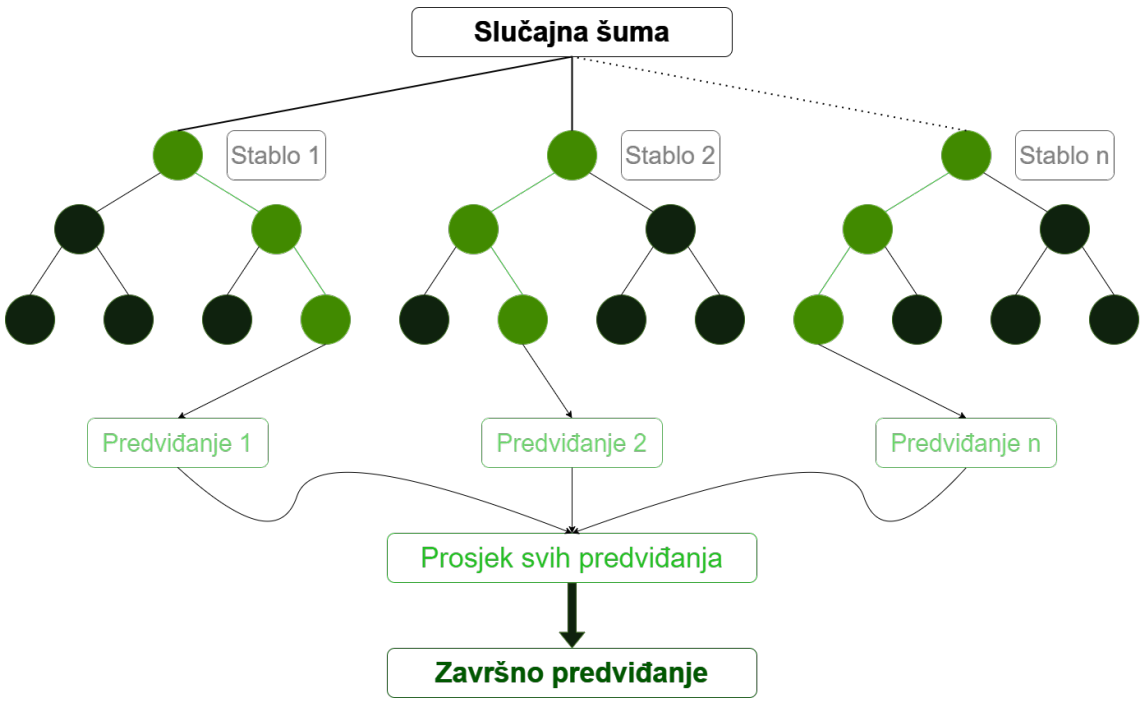
```
1    gbm_model <- gbm(
2        formula = TEC ~ Bx + By + Bz,
3        data = training_data,
4        distribution = "gaussian",
5        n.trees = 100,
6        interaction.depth = 3,
7        shrinkage = 0.01,
8        cv.folds = 5
9       )
```

Isječak koda 4.3 Funkcija gradijentnog pojačavanja u R-u

4.3.4 Slučajna šuma

Slučajna šuma predstavlja grupnu metodu koja kombinira više stabala odluke za smanjenje varijabilnosti modela i poboljšanje točnosti predviđanja. U slučaju klasi-

fikacije konačno predviđanje temelji se na većinskom glasanju svih stabala, dok se u regresiji konačno predviđanje temelji na prosjeku predviđanja svih stabala [31], što je u pojednostavljenom obliku prikazano na slici 4.7.



Slika 4.7 Pojednostavljeni prikaz načina rada slučajne šume

Za osiguranje robusnosti modela ključne su prvenstveno dvije tehnike:

- *Bootstrap* uzorkovanje - za svako stablo unutar šume model odabire nasumični uzorak s ponavljanjem iz skupa podataka za treniranje modela
- Slučajan odabir značajki - kod svake podjele čvora u stablu se slučajno odabire podskup značajki na koje se čvor može podijeliti, što osigurava jedinstvenost svakog stabla i smanjuje korelaciju među njima

Konačno predviđanje modela $\hat{y}$ za određeni podatak dobiva se na temelju prosjeka svih predviđanja kako je vidljivo u jednadžbi 4.6, gdje je N broj stabala u slučajnoj

šumi a $\hat{y}_i$ predviđanje pojedinog stabla.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i \tag{4.6}$$

Modelu i funkcijama slučajne šume u R-u se pristupa korištenjem paketa *randomForest* (korištena verzija 4.7-1.1) [32], a osnovna varijanta funkcije korištene za stvaranje modela kao parametre prima vrijednosti prediktora, ciljne vrijednosti te broj stabala koji će biti korišteni u modelu, kako je prikazano na isječku koda 4.4.

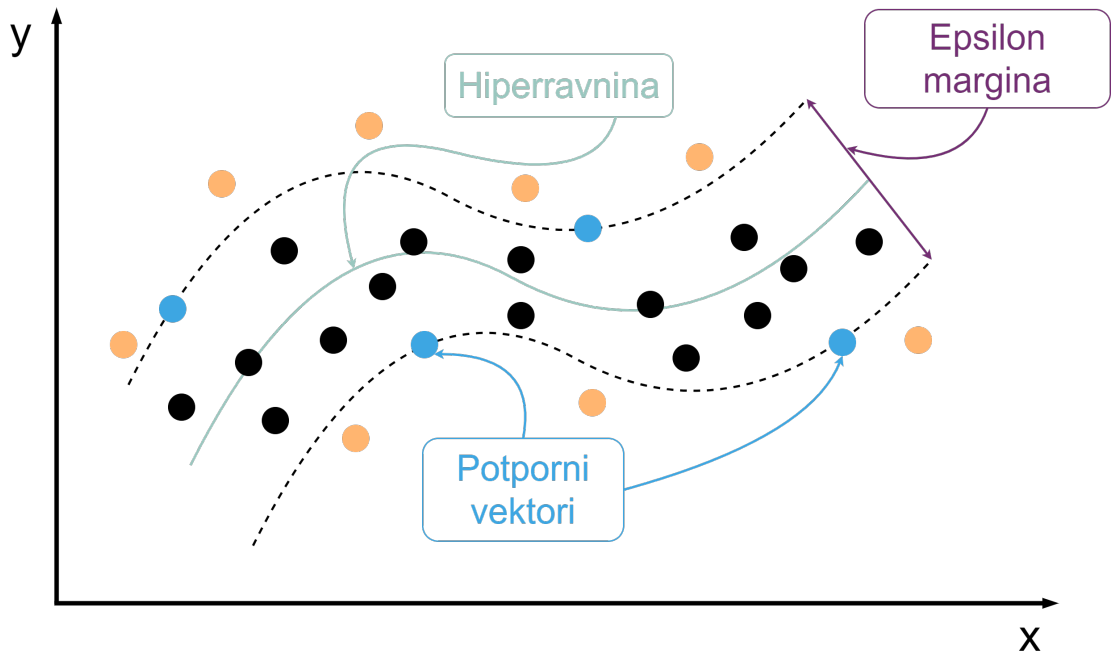
```
1    random_forest_model <- randomForest(
2        x = train_data_predictors,
3        y = training_data$TEC,
4        ntree = 100
5    )
```

Isječak koda 4.4 Funkcija slučajne šume u R-u

4.3.5 Stroj potpornih vektora

Stroj potpornih vektora metoda je strojnog učenja koja u slučaju regresije, kada se naziva potporna vektorska regresija (eng. Support Vector Regression (SVR)), koristi koncept marginalizacije i potpornih vektora te pokušava optimalno uklopiti najbolju liniju, tj. hiperravninu, unutar određene granice tolerancije, epsilon margine, koja minimizira pogreške između predviđanja i stvarnih vrijednosti tražene varijable. Cilj je osigurati da se predviđene vrijednosti nalaze unutar dane granice, dok se ona predviđanja koja su van nje kažnjavaju i doprinose funkciji gubitka modela [33]. Potporni vektori podatkovne su točke najbliže zadanoj regresijskoj liniji i služe za

definiranje položaja i orijentacije regresijske funkcije. Opisani način rada vidljiv je u pojednostavljenom obliku na slici 4.8.



Slika 4.8 Pojednostavljeni prikaz načina rada slučajne šume

Matematički izraz za predviđanje vrijednosti pomoću stroja potpornih vektora prikazan je u jednadžbi 4.7:

$$\hat{y} = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{4.7}$$

gdje je:

- α_i - težina dodijeljena svakom potpornom vektoru
- $K(x_i, x)$ - funkcija jezgre koja mjeri sličnosti između podatka x_i i trenutnog ulaza x
- b - pomak (eng. Bias)

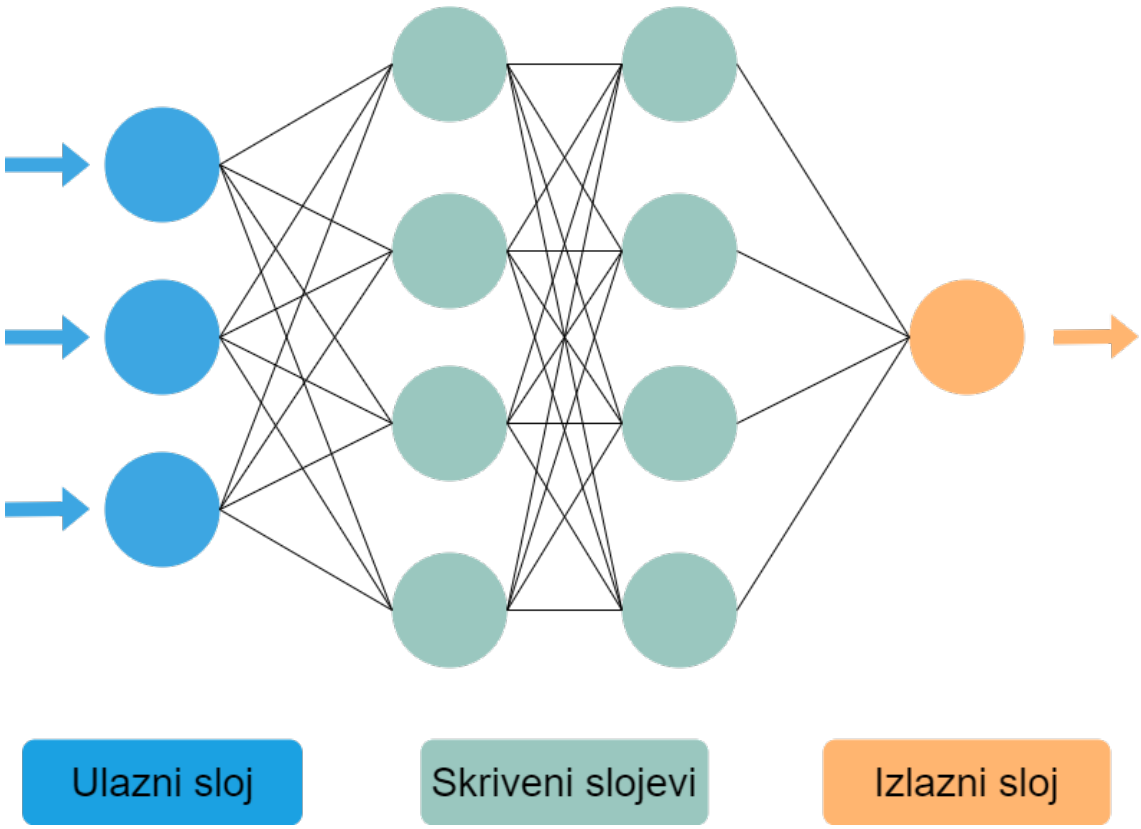
R paket *e1071* [34] koristi se za pristup modelu stroja potpornih vektora, a primjer korištenja funkcije *svm*() koja stvara model vidljiv je na isječku koda 4.5. U nju se redom šalju parametri: jednadžba u kojoj su određeni prediktori i ciljna vrijednost, skup podataka koji je korišten za treniranje modela, funkcija jezgre koja definira način transformiranja ulaznih podataka te naposljetku vrsta modela. Korištena je radijalna funkcija jezgre koja modelu omogućuje nošenje s nelinearnim odnosima među podatcima, a vrsta modela je epsilon-regresija kojom se postiže ravnoteža između točnosti predviđanja i otpornosti na netipične vrijednosti.

```
1    svm_model <- svm(
2        formula = TEC ~ Bx + By + Bz,
3        data = training_data,
4        kernel = "radial",
5        type = "eps-regression"
6        )
```

Isječak koda 4.5 Funkcija stroja potpornih vektora u R-u

4.3.6 Neuronska mreža

Neuronske mreže sastoje se od slojeva umjetnih neurona, organiziranih u ulazni sloj, skrivene slojeve i izlazni sloj, gdje svaki neuron prima ulazne podatke nad kojima primjenjuje aktivacijske funkcije i određene težine te dobivene izlazne podatke prenosi dalje kroz mrežu. Prikaz jednog osnovnog modela duboke neuronske mreže, mreže koja ima više od jednog skrivenog sloja za bolje učenje složenih uzoraka i nelinearnih pristupa podatcima, vidljiv je na slici 4.9. Glavni koncept treniranja je pronalazak optimalnih težina koje minimiziraju razliku između predviđenih i stvarnih vrijednosti.



Slika 4.9 Pojednostavljeni prikaz dijagrama neuronske mreže

Duboka neuronska mreža razvijena u radu izvedena je uz pomoć R paketa *keras* [35] verzije 2.15.0, koji omogućuje intuitivnu i brzu izgradnju neke od podržanih, različitih vrsta neuronskih mreža (konvolucijske, rekurzivne, kombinirane itd.), te R paketa *tensorflow* [36] 2.16.0., koji pruža učinkovitu numeričku obradu, paralelizaciju i optimizaciju modela. Razvijena neuronska mreža definirana je kao sekvencijalni model u kojem se podatci kroz mrežu prenose linearno s tri potpuno povezana (eng. Dense) sloja, što znači da je svaki neuron unutar nekog sloja povezan sa svim neuronima iz njemu prethodnog sloja, a primjer stvaranja modela neuronske mreže nalazi se na isječku koda 4.6. Izrađeni model sastoji se od 4 sloja: ulaznog sloja koji prima komponente magnetskog polja Bx, By i Bz kao ulazne parametre, dva skrivena sloja

od kojih svaki ima 64 neurona i aktivacijsku funkciju ReLU (eng. Rectified Linear Unit) koja unosi nelinearnost u model i time omogućuje modelu da nauči složene obrasce iz podataka te izlazni sloj koji se, s obzirom da je riječ o regresiji, sastoji od jednog neurona kojemu je cilj predviđanje jedne kontinuirane vrijednosti; TEC-a.

```
model <- keras_model_sequential() %>%
    layer_dense(
        units = 64,
        activation = 'relu',
        input_shape = ncol(train_data)
    ) %>%
    layer_dense(
        units = 64,
        activation = 'relu'
    ) %>%
    layer_dense(units = 1)
```

Isječak koda 4.6 Izrada modela neuronske mreže u R-u

Konstruirani model potrebno je kompilirati kako je prikazano na isječku koda 4.7, gdje se još kao parametre predaje funkcija gubitka u kojoj se koristi srednja kvadratna pogreška (eng. Mean Square Error (MSE)) koju se cilja minimizirati tijekom treniranja modela, optimizacijski algoritam RMSprop koji automatski regulira stope učenja i stabilizira treniranje te metrika srednje apsolutne pogreške (eng. Mean Absolute Error (MAE)) između predviđanja i stvarnih vrijednosti koja se koristi za vrednovanje modela.

```
model %>% compile(
    loss = 'mse',
    optimizer = optimizer_rmsprop(),
    metrics = c('mae')
)
```

Isječak koda 4.7 Kompiliranje modela neuronske mreže u R-u

Za kraj je potrebno istrenirati sačinjeni model, a način izvedbe istog vidljiv je u funkciji na isječku koda 4.8 unutar koje su definirani podatci za treniranje modela, broj epoha kojim je definiran broj prolaska modela kroz cijeli skup tih podataka, gdje se svaki prolaskom uči kako bolje prilagoditi težine, broj uzoraka koje model obrađuje prije ažuriranja težina (manji broj rezultira češćim ažuriranjima, ali može biti osjetljiviji na šum unutar podataka) te postotak kojim je određeno koliki postotak predanih podataka za treniranje će biti korišten za validaciju modela, kako bi se izbjegla pretjerana prilagodba modela.

```
neural_network_model <- model %>% fit(
    train_data,
    train_labels,
    epochs = 50,
    batch_size = 32,
    validation_split = 0.2
)
```

Isječak koda 4.8 Treniranje modela neuronske mreže u R-u

Predviđanje pojedinog sloja neuronske mreže prikazano je jednadžbom 4.8:

$$\hat{y} = f \cdot (Wx + b) \tag{4.8}$$

gdje:

- W predstavlja težine
- x predstavlja ulaznu vrijednost (prediktore Bx, By i Bz)
- b predstavlja pomak (bias)
- f predstavlja aktivacijsku funkciju (ReLU)

Općenito svaki sloj može biti predstavljen kao linearna kombinacija ulaznih značajki popraćena nelinearnom aktivacijskom funkcijom. Prolaskom kroz slojeve mreže, dodaju se složenije transformacije na temelju težina i pomaka unutar svakog sloja te je konačni matematički izraz jednak onom prikazanom u jednadžbi 4.9.

$$\hat{y} = W_3 \cdot f_2 \left(W_2 \cdot f_1 \left(W_1 x + b_1\right) + b_2\right) + b_3 \tag{4.9}$$

Gdje su:

- W_1, W_2 i W_3 matrice težina za prvi, drugi i treći sloj
- b_1, b_2 i b_3 pomaci za svaki sloj
- f_1 i f_2 aktivacijsku funkcije u skrivenim slojevima

4.4 Vrednovanje modela

Razvijeni modeli vrednuju se na temelju tri ključna kriterija:

1. Predviđeno-izmjereni dijagram (eng. Prediction-Observation Diagram) zajedno s dijagramom kumulativne distribucije vjerojatnosti (eng. Predicted vs Predicted Probability Diagram)

2. Korijen srednje kvadratne pogreške RMSE
3. Prilagođeni koeficijent determinacije (eng. Adjusted R-Squared)

Predviđeno-izmjereni dijagram prikazuje usporedbu predviđenih i stvarnih, izmjerenih vrijednosti, gdje je idealna linija $x = y$ indikator savršenog predviđanja; što su joj točke bliže, to je model bolji. Dijagram kumulativne distribucije vjerojatnosti prikazuje kumulativne frekvencije predviđanja u odnosu na izmjerene, koristeći se također istim principom idealne linije. Predviđeno-izmjereni dijagram pruža uvid u to koliko se dobro pojedinačna predviđanja podudaraju s izmjerenim vrijednostima TEC-a, dok dijagram kumulativne distribucije pomaže prikazati slijede li predviđene TEC vrijednosti modela jednaku distribuciju kao izmjerene vrijednosti.

Korijen srednje kvadratne pogreške je metrika koja mjeri prosječnu veličinu greške između predviđenih i izmjerenih vrijednosti, a izračun je prikazan u jednadžbi 4.10, gdje n predstavlja broj uzoraka, y_i stvarne vrijednosti i $\hat{y}_i$ predviđanja modela.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.10}$$

Prilagođeni koeficijent determinacije dobiva se iz običnog koeficijenta determinacije koji nam govori koliko dobro nezavisne varijable (prediktori) predviđaju zavisnu varijablu. Za razliku od njega, prilagođeni koeficijent determinacije uvodi sankcije kod korištenja pretjeranog broja prediktora, čime se izbjegava pretjerana prilagodba modela. Izračun prilagođenog koeficijenta determinacije prikazan je u jednadžbi 4.11:

$$\text{Prilagođeni } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1}\right) \tag{4.11}$$

gdje je:

- R^2 - obični koeficijent determinacije

- n - broj uzoraka
- p - broj prediktora

Osim ova glavna tri kriterija, kako bi se dobio bolji uvid u kvalitetu svakog pojedinog modela iz još nekoliko različitih perspektiva, za svaki model provjerena su još tri metrička pokazatelja: srednja apsolutna pogreška (MAE), maksimalna vrijednost reziduala i vrijeme treniranja modela izraženo u sekundama. Srednja apsolutna pogreška mjeri apsolutnu razliku između svih izmjerenih i predviđenih vrijednosti te je manje osjetljiva na ekstremne, netipične vrijednosti nego li RMSE. Maksimalna vrijednost reziduala predstavlja najveću pojedinačnu grešku između svih predviđenih vrijednosti modela, odnosno najveće odstupanje predviđene vrijednosti od izmjerene.

Procjena izvedbe modela temeljena na ovim kriterijima primijenjena je na šest razvijenih prognostičkih modela ionosferskog kašnjenja kako bi ih se moglo međusobno usporediti.

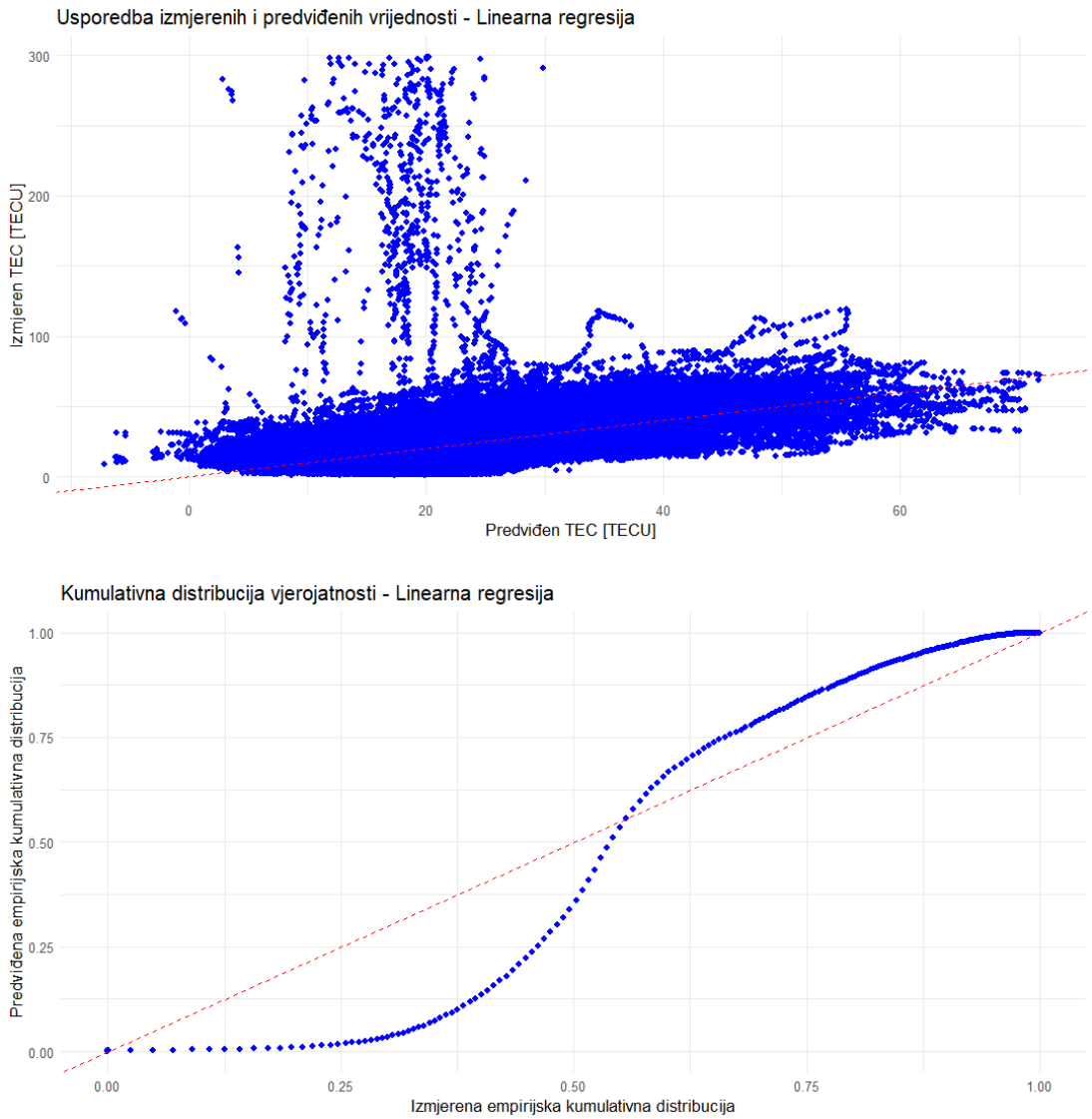
Poglavlje 5

Rezultati istraživanja

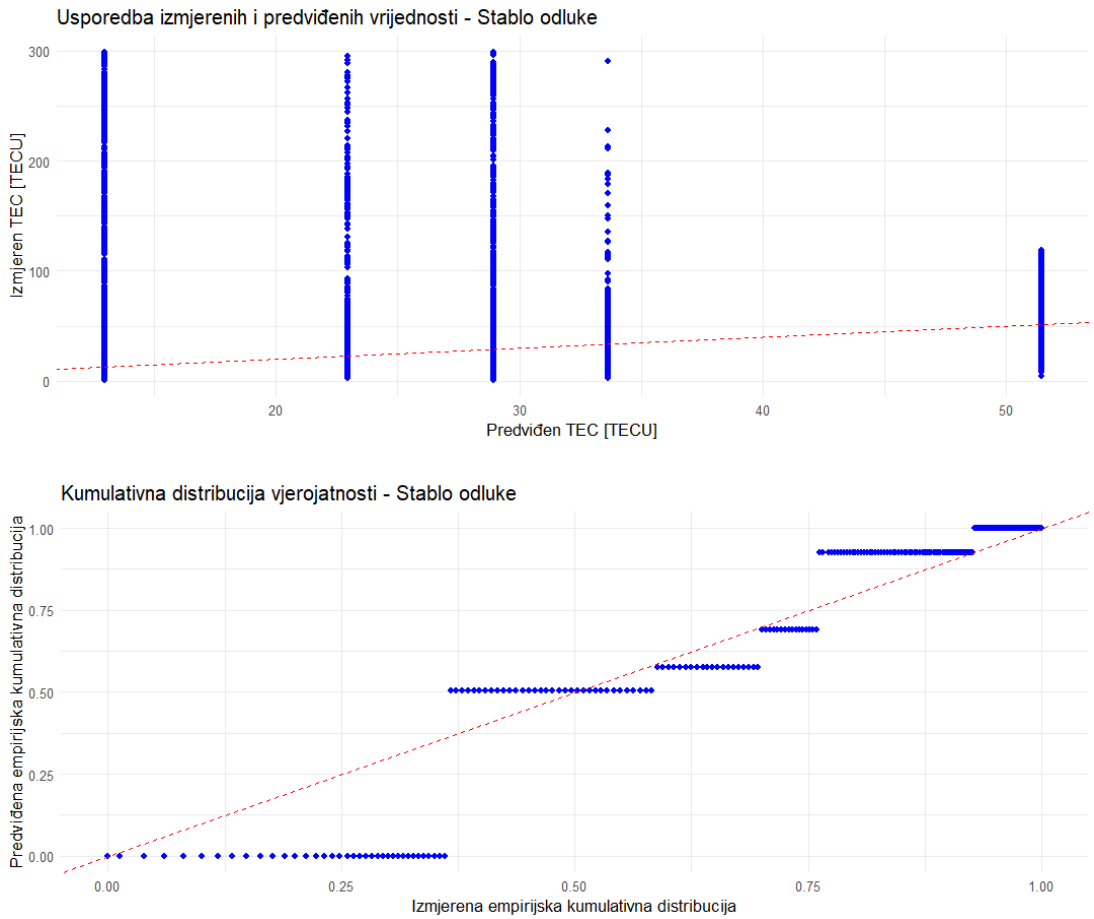
Sukladno opisanoj metodologiji, metodama strojnog učenja razvijeno je 6 prognostičkih modela GPS ionosferskog kašnjenja s komponentama geomagnetskog polja kao prediktorima. Za svaki od razvijenih modela u tablici 5.1 zapisane su dobivene vrijednosti metričkih pokazatelja, redom; srednja apsolutna pogreška MAE, korijen srednje kvadratne pogreške RMSE, maksimalni rezidual, prilagođeni koeficijent determinacije i vrijeme razvoja modela, dok su njihovi predviđeno-izmjereni dijagrami zajedno s dijagramima kumulativne distribucije vjerojatnosti prikazani na slikama 5.1 do 5.6 gdje svaka plava točka predstavlja jedan par izmjerene i predviđene vrijednosti. Na slici 5.7 su grafički prikazane vrijednosti korijena srednje kvadratne pogreške modela (označene plavom bojom) i vrijednosti prilagođenog koeficijenta determinacije modela (označene ružičastom bojom).

Tablica 5.1 Usporedba rezultata razvijenih modela

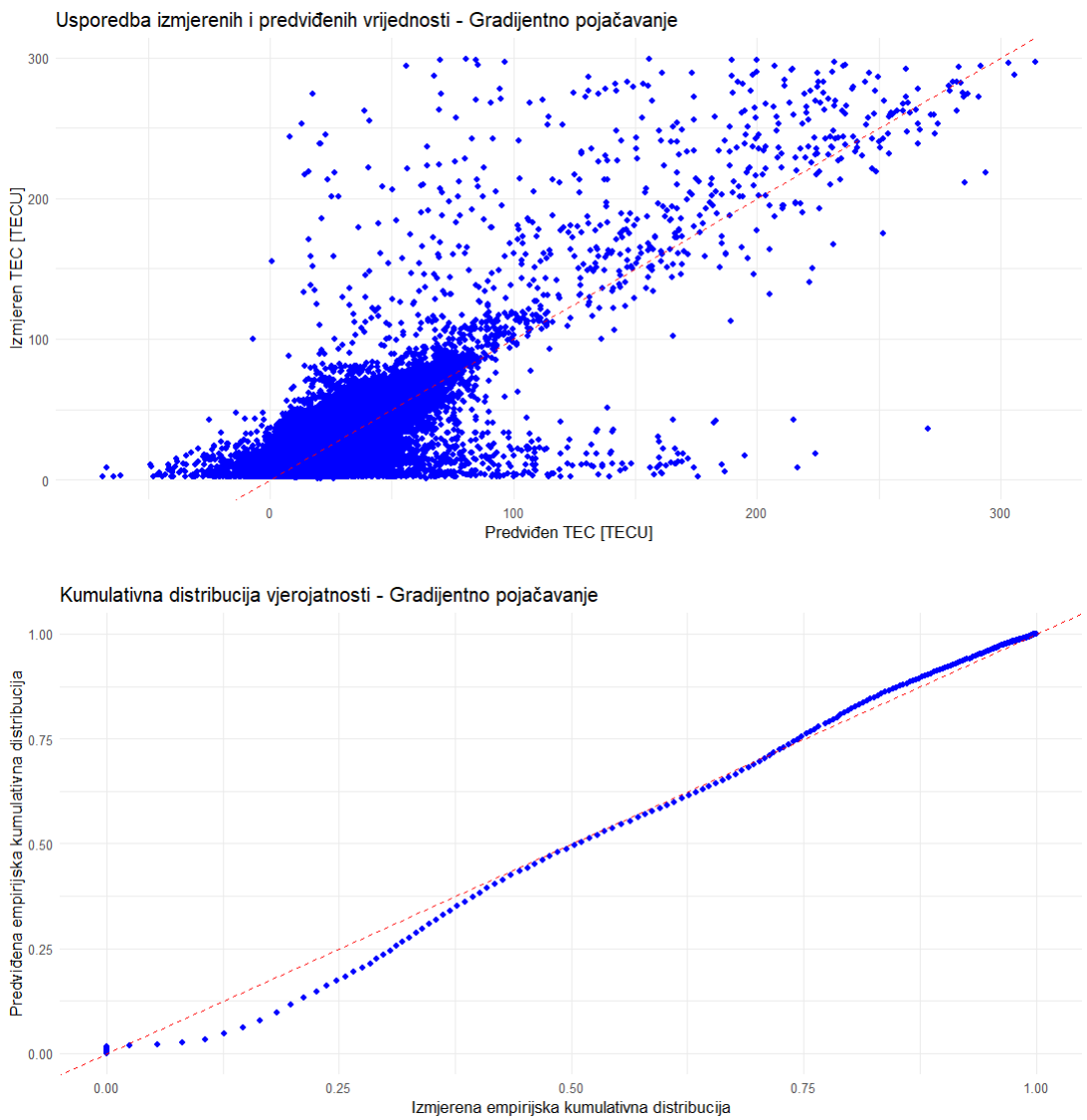
Model	MAE $[TECU]$	**RMSE** $[TECU]$	Maksimalni rezidual $[TECU]$	**Prilagođeni** R^2 [%]	Vrijeme razvoja $[s]$
Linearna regresija	12,12412	19,83721	294,486	0,1977	0,07
Stablo odluke	10,55777	18,7592	286,071	0,2805	3,04
Gradijentno pojačavanje	6,688199	11,59750	216,702	0,8427	3359
Slučajna šuma	4,856270	10,52184	288,358	0,7459	4410
Stroj potpornih vektora	8,120794	17,91729	295,032	0,3499	61531
Neuronska mreža	8,566005	17,28053	283,759	0,3953	544



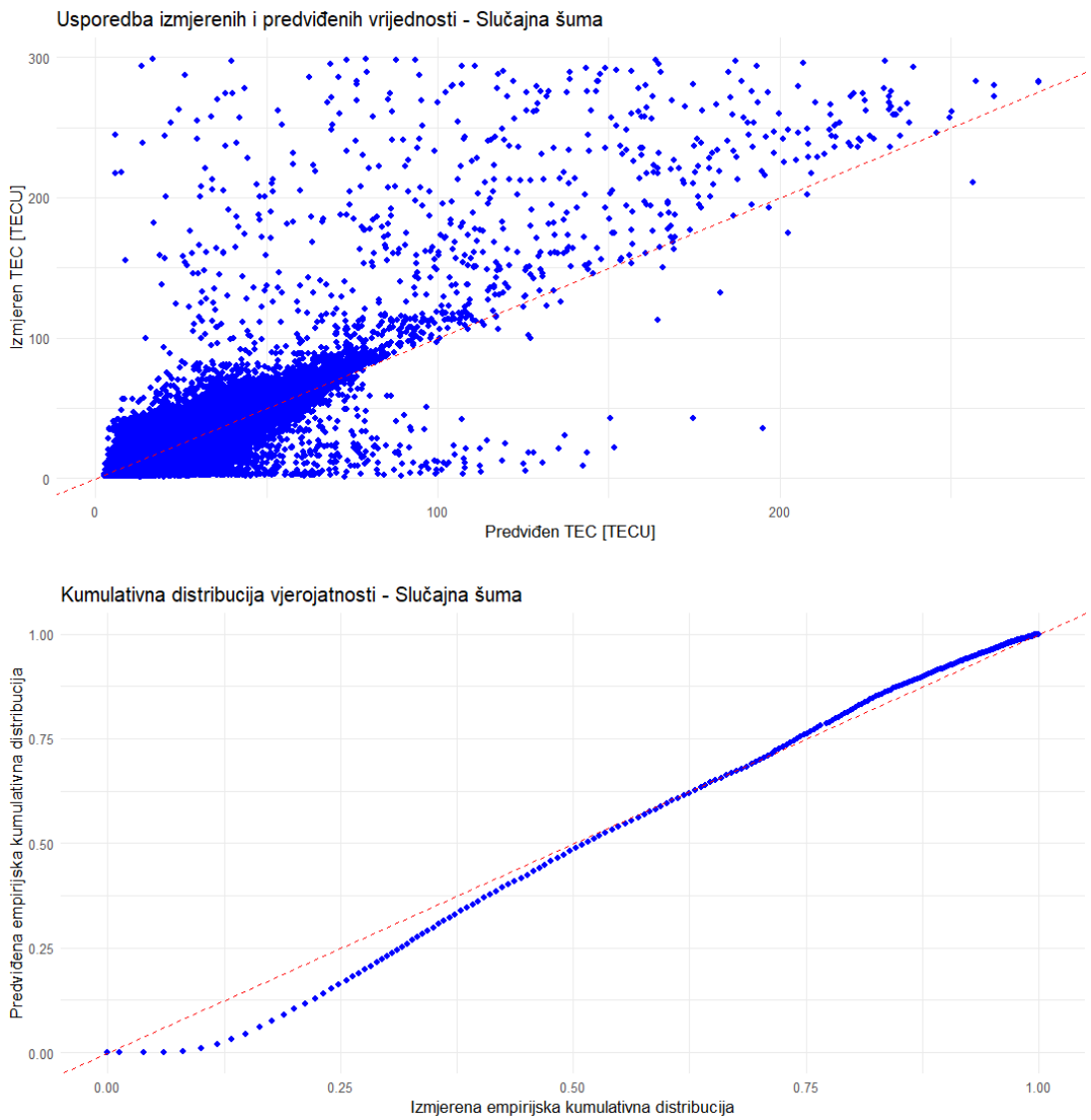
Slika 5.1 Predviđeno-izmjereni dijagram i dijagram kumulativne distribucije linearne regresije



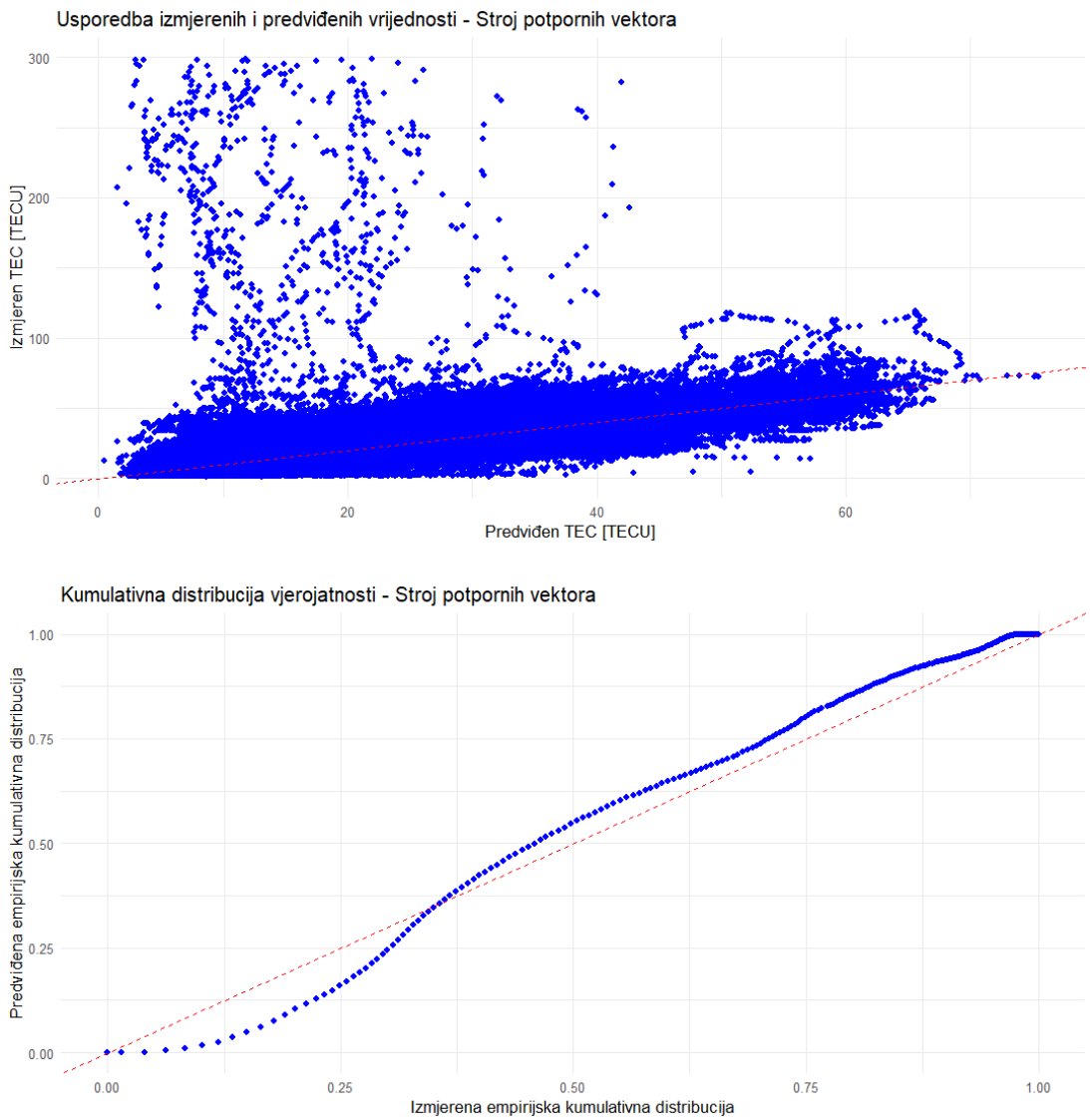
Slika 5.2 Predviđeno-izmjereni dijagram i dijagram kumulativne distribucije stabla odluke



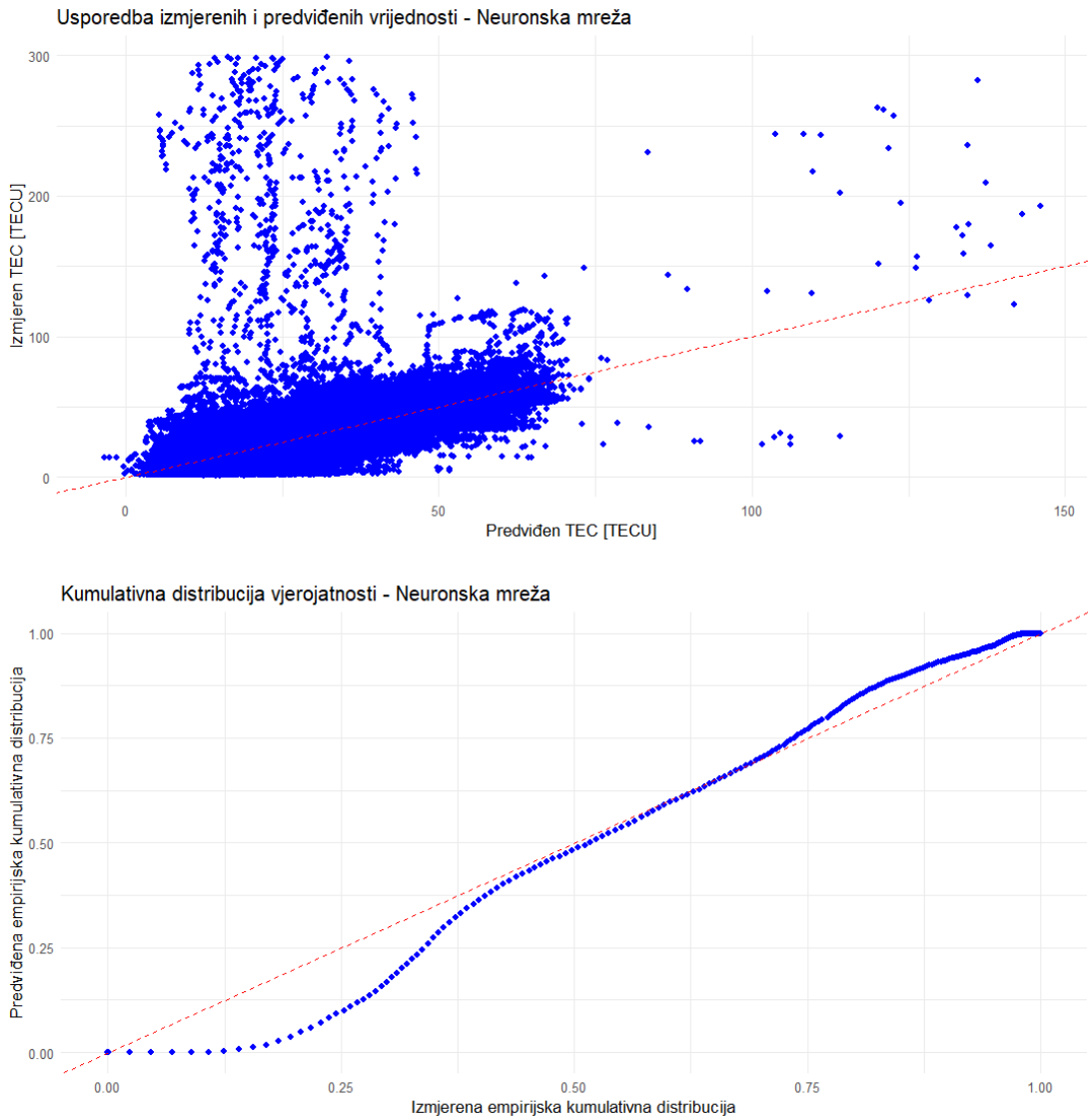
Slika 5.3 Predviđeno-izmjereni dijagram i dijagram kumulativne distribucije gradijentnog pojačavanja



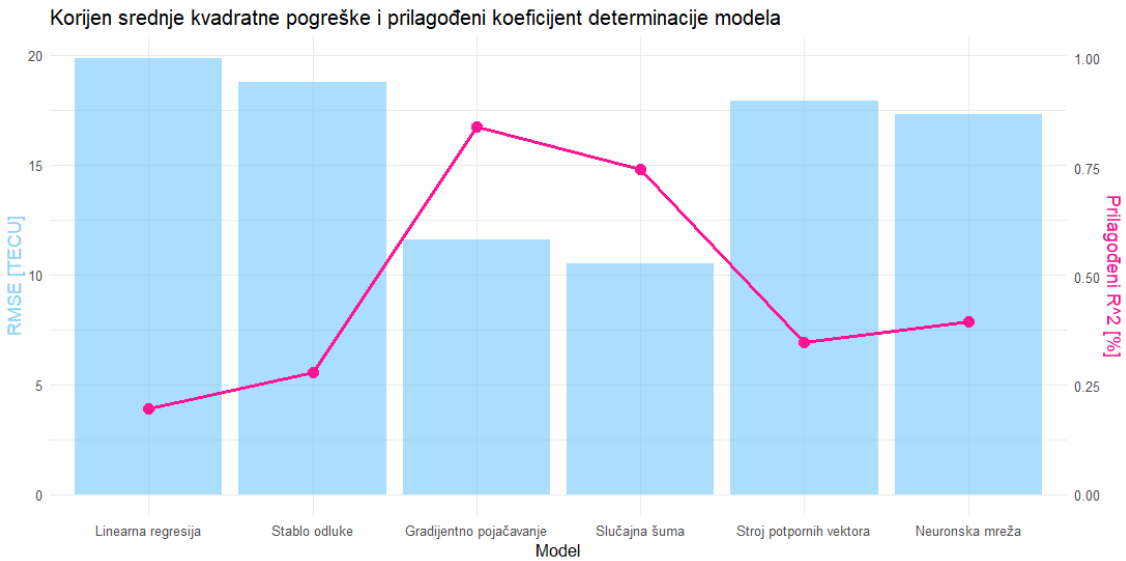
Slika 5.4 Predviđeno-izmjereni dijagram i dijagram kumulativne distribucije slučajne šume



Slika 5.5 Predviđeno-izmjereni dijagram i dijagram kumulativne distribucije stroja potpornih vektora



Slika 5.6 Predviđeno-izmjereni dijagram i dijagram kumulativne distribucije neuronske mreže



Slika 5.7 Vrijednosti korijena srednje kvadratne pogreške i prilagođenog koeficijenta determinacije razvijenih modela

Poglavlje 6

Interpretacija

U nastavku su za svaki model interpretirani predstavljeni rezultati, komentirane su dobivene vrijednosti te navedene osnovne prednosti i mane svakog modela s obzirom na dobivene rezultate. Na kraju su međusobno uspoređeni, s fokusom na one najbolje, a predstavljene su i ideje na temelju kojih bi se modeli i sam izračun ionosferskog kašnjenja mogli poboljšati.

6.1 Linearna regresija

Prognostički model ionosferskog kašnjenja izveden uz pomoć linearne regresije daje procjene ionosferskog kašnjenja u rasponu od -7,1875 do 71,7324 TEC-a kako je vidljivo iz njegovog predviđeno-izmjerenog dijagrama 5.1. Velika raspršenost točaka, pogotovo u nižim vrijednostima do 30 TEC-a, znak je poteškoća kod preciznog predviđanja, posebice uočljivo kod većih vrijednosti TEC-a. Opisani ishod je očekivan budući da je model osjetljiv na netipične vrijednosti, a njihovo postojanje unutar podataka potvrđuje i izrazito velika vrijednost maksimalnog reziduala, kao i vrijednosti srednje apsolutne pogreške MAE i korijena srednje kvadratne pogreške RMSE, kada se usporede sa srednjom vrijednosti ciljne varijable TEC koja iznosi 23,17844

TECU. Dijagram kumulativne distribucije vjerojatnosti linearne regresije ukazuje kako je distribucija predviđenih vrijednosti vidno različita od distribucije izmjerenih; model podcjenjuje distribuciju nižih vrijednosti TEC-a u malo više od 50% prvotnih vrijednosti izmjerene kumulativne distribucije, a potom precjenjuje distribuciju viših. Iz prilagođenog koeficijenta determinacije jasno je da model objašnjava malo manje od 20% varijabilnosti ciljne varijable, odnosno ukazuje kako linearna regresija nije adekvatna za modeliranje ovih podataka, jer je jedina pozitivna karakteristika modela brzina treniranja i predviđanja.

Model linearne regresije je brz, računarski učinkovit te jednostavan za izvršavanje i tumačenje. Pretpostavlja linearnost između prediktora i ciljne vrijednosti, što u složenijim fenomenima, kao što je ionosferski utjecaj, često nije slučaj. Bez obzira na činjenicu da je model osjetljiv na multikolinearnost između prediktora, koja je unutar korištenih podataka samo slaba do umjerena, te doradu modela prema međusobnoj korelaciji podataka temeljenoj na već prikazanom korelogramu, izvedba modela je ograničena jer ne postoje dovoljno snažni linearni odnosi među podatcima, što u konačnici rezultira visokom razinom pogreške.

6.2 Stablo odluke

Predviđanja koja daje model izrađen kao stablo odluke razlikuju se od svih drugih modela jer se radi o nekoliko diskretnih razina predviđenih vrijednosti zbog prirode rada algoritma, prikazanih na predviđeno-izmjerenom dijagramu 5.2. Predviđene vrijednosti su redom 12,92862, 22,93331, 28,94894, 33,64144 i 51,45559 TEC-a, samo pet vrijednosti predviđeno nakon testiranja modela nad 101 977 instanci podataka za testiranje. Usprkos pretjeranoj prilagodbi podataka, metrike poput srednje apsolutne vrijednosti, korijenu srednje kvadratne pogreške, vrijednosti maksimalnog reziduala te prilagođenog koeficijenta determinacije bolje su u odnosu na iste metrike modela

razvijenog pomoću linearne regresije, dok je vrijeme treniranja malo slabije no i dalje izuzetno brzo. Iz dijagrama kumulativne distribucije vjerojatnosti uočljivo je kako model do 75% vrijednosti na apscisi gotovo konstantno podcjenjuje izmjerenu kumulativnu distribuciju i predviđa brojčano manje vrijednosti nego li ih zaista ima u izmjerenim podatcima.

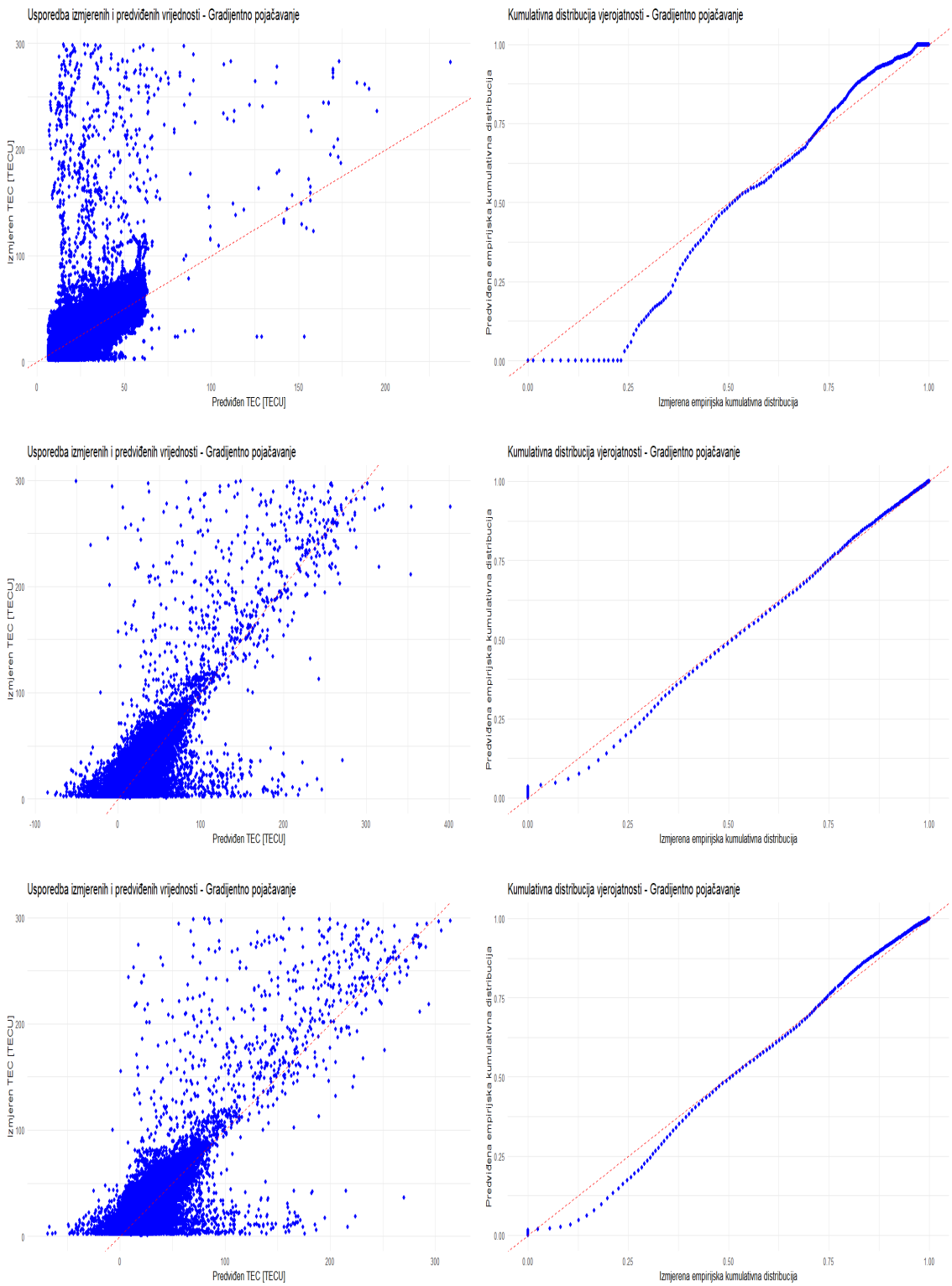
Glavne prednosti modela stabla odluke su brzina, jednostavnost tumačenja i sposobnost modeliranja nelinearnih, složenih odnosa među varijablama. Diskretna priroda predviđenih vrijednosti proizlazi iz načina na koji stablo odluke dijeli podatke u odvojene skupine, što je prednost kada se radi sa izrazito heterogenim podatcima, no dovodi do neispravnih rezultata kada se radi o kontinuiranim i kompleksnim varijablama poput TEC-a. Model je sklon pretjeranoj prilagodbi podataka u slučajevima kad nema odgovarajućeg obrezivanja (eng. Pruning) i kad je dopuštena prevelika dubina ili kompleksnost stabla. Nestabilnost modela također je problem, budući da čak i male promjene u podatcima mogu značajno izmijeniti odluke unutar stabla, a ona se rješava korištenjem više stabala odluke u grupnim, ansambl metodama.

6.3 Gradijentno pojačavanje

Model izrađen koristeći metode gradijentnog pojačavanja prema predviđeno-izmjerenom dijagramu 5.3 daje relativno dobra predviđanja vrijednosti do 90 TEC-a, ali daje i nekolicinu negativnih predviđanja, nešto što nije očekivano na temelju skupa podataka za treniranje i stvarnih vrijednosti TEC-a u ionosferi, koje nikada ne mogu biti negative. Na dijagramu kumulativne distribucije vjerojatnosti primjetno je kako model, osim u prvih 35% vrijednosti izmjerene kumulativne distribucije gdje su vrijednosti podcijenjene, većinski prati distribuciju stvarnih podataka. Dobivene metrike ukazuju kako je model prilično dobar u usporedbi s ostalim razvijenim modelima, pogotovo prema metrici prilagođenog koeficijenta determinacije po kojoj objaš-

njava gotovo 85% varijabilnosti u stvarnim vrijednostima TEC-a, ali je po pitanju vremena potrebnog za razvoj modela među lošijima.

Na dobivene vrijednosti najveći utjecaj imaju hiperparametri, parametri koji se ne uče izravno iz podataka predanih modelu već ih je potrebno postaviti prije pokretanja obuke modela, a taj utjecaj vidljiv je na rezultatima prikazanim na slici 6.1 gdje su uspoređeni predviđeno-izmjereni dijagrami i dijagrami kumulativne distribucije vjerojatnosti triju modela nastalim na temelju različitih vrijednosti hiperparametara prikazanih u tablici 6.1, zajedno s glavnim kriterijima vrednovanja i vremenom potrebnim za njihov razvoj.



Slika 6.1 Usporedba dijagrama tri različita modela gradijentnog pojačavanja

Tablica 6.1 Usporedba rezultata razvijenih modela gradijentnog pojačavanja

Model	Broj stabala	Dubina stabla	Brzina učenja	Broj presavijanja	RMSE $[TECU]$	Prilagođeni R^2 [%]	Vrijeme $[s]$
1.	300	9	0,03	6	16,62292	0,4452255	201,7
2.	4000	20	0,6	16	12,21439	0,9004171	13896
Finalni	2000	16	0,4	12	11,59750	0,8427412	3359

Gradijentno pojačavanje kao okosnica razvijenog modela osigurava robusnost prema nelinearnosti, sposobnost generalizacije zahvaljujući unakrsnim provjerama valjanosti te fleksibilnost modela zbog mogućnosti podešavanja hiperparametara kao što su broj stabala, dubina stabala i brzina učenja. Za učinkovit i točan rad modela potrebna je velika količina podataka jer na manjim skupovima može dolaziti do slabijih izvedba i većih varijabilnost, a velika osjetljivost na hiperparametre igra značajnu ulogu u izvedbi modela pa pravilno podešavanje istih može biti zahtjevno i iziskivati puno vremena. Model je složen, stoga je njegovo tumačenje manje intuitivno, a zbog iterativnog procesa dodavanja podmodela (stabala) treniranje može biti sporo, posebice kada se koristi veliki broj stabala ili je potrebno analizirati veliki skup podataka, što zahtjeva i znatno više računalnih resursa.

6.4 Slučajna šuma

Razvijeni model metodom slučajne šume procjenjuje vrijednosti ionosferskog kašnjenja u rasponu 2,4 - 275,7 TEC-a, no kako je vidljivo na predviđeno-izmjerenom dijagramu 5.4, precizna predviđanja donose se do približno 85 TEC-a. Premda je većina predviđanja koja je udaljena od idealne $x = y$ linije dana za niže vrijednosti TEC-a, model i dalje podcjenjuje početnih 35% vrijednosti izmjerene kumulativne distribucije (prikazano na dijagramu kumulativne distribucije vjerojatnosti), ali nakon tog praga dobro je praćena distribucija stvarnih podataka. Male vrijednosti sred-

nje apsolutne pogreške i korijena srednje kvadratne pogreške, u usporedbi s ostalim razvijenim modelima, odražavaju sposobnost modela da dovoljno uspješno predviđa ionosfersko kašnjenje. Vrijednost maksimalnog reziduala sugerira kako za netipične vrijednosti model može imati izuzetno pogrešna predviđanja, a vrijeme potrebno za razvoj među najlošijima je od razvijenih modela.

Sve dobivene vrijednosti odnose se na model za čije je razvijanje korištena slučajna šuma sa 100 stabala. Razvijeni su još modeli koji imaju manje i koji imaju više stabala, no rezultati su bili ili lošiji ili minimalno bolji po glavnim kriterijima od predstavljenog modela, nauštrb neke od ostalih metrika. Usporedba tih modela prikazana je u tablici 6.2.

Tablica 6.2 Usporedba rezultata razvijenih modela slučajne šume

Model	Broj stabala	MAE $[TECU]$	RMSE $[TECU]$	Maksimalni rezidual $[TECU]$	Prilagođeni R^2 [%]	Vrijeme razvoja $[s]$
1.	50	4,8829	10,584	287,173	0,7229	2096
2.	300	4,8289	10,465	287,757	0,7656	14015
Finalni	100	4,8562	10,522	288,358	0,7459	4410

Grupna metoda sastavljena od većeg broja stabala odluke, slučajna šuma, osigurava stabilnost modela u kojem se koristi, pruža veću otpornost modela na šum i štiti ga od pretjerane prilagodbe. Omogućuje rad s velikim i kompleksnim skupovima podataka, a ne zahtijeva pretpostavke o distribuciji podataka, što model čini vrlo fleksibilnim. Kompleksnost modela i njegovo tumačenje te osjetljivost na iznimno neuravnotežene podatke neke su od poteškoće s kojima se model razvijen korištenjem slučajne šume susreće. Ovisno o broju stabala u modelu, treniranje može zahtijevati značajni vremenski period, uz kojeg se veže i korištenje većeg broja računalnih resursa, pogotovo kad je riječ o većim i/ili kompleksnijim skupovima podataka.

6.5 Stroj potpornih vektora

Model razvijen korištenjem stroja potpornih vektora, odnosno potpornom vektorskom regresijom, predviđa vrijednosti do maksimalno 75,47982 TEC-a, iako ih je velika većina raspoređeno do 65 TEC-a, a najveća raspršenost točaka događa se kod predviđenih vrijednosti do 30 TEC-a, gdje su za neuobičajeno visoke izmjerene vrijednosti ionosferskog kašnjenja pogrešno predviđene niske vrijednosti, što je uočljivo na predviđeno-izmjerenom dijagramu 5.5. Unatoč tome, prema dijagramu kumulativne distribucije vjerojatnosti model i dalje unutar prvih 30% vrijednosti izmjerene kumulativne distribucije predviđa brojčano manje vrijednosti nego li ih je izmjereno. Prema izmjerenim metrikama model se posebno ne ističe niti pozitivno niti negativno u usporedbi s drugim razvijenim modelima, osim po metrici vremena razvoja. Vrijednošću od 61 531 sekunde, odnosno nešto više od 17 sati, razvijanje modela korištenjem stroja potpornih vektora gotovo je 14 puta sporije nego li kod idućeg najsporije razvijenog modela; slučajne šume, koja je bolja i po svakom ostalom kriteriju.

Stroj potpornih vektora izrađuje model koji je fleksibilan i ima sposobnost modeliranja nelinearnih odnosa između podataka zahvaljujući raznim jezgrenim funkcijama, poput polinomne, sigmoidne ili korištene radijalne, koje omogućuju prilagodbu kompliciranijim podatcima i otkrivanje nelinearnih obrazaca među njima. Model se odlikuje i robusnošću prema netipičnim podatcima, a posebice je učinkovit kada su u pitanju manji skupovi podataka. Računalna složenost modela u ovom je slučaju najveći problem jer raste zajedno s porastom broja uzoraka u podatcima, posljedično stvarajući i problem skalabilnosti modela, što se odražava i na vremenu potrebnim za treniranje. Zbog navedenog je i tumačenje modela kompliciranо, a pri korištenju

posebnu pozornost treba obratiti na izbor i detaljnije podešavanje jezgrene funkcije, kao i skaliranje podataka ako je potrebno, da bi se izbjegla dominacije jedne varijable nad drugima.

6.6 Neuronska mreža

Četveroslojna duboka neuronska mreža predviđa vrijednosti ionosferskog kašnjenja u rasponu od -3,4482 do 146,0491 TEC-a, uz najveću raspršenost točaka kod nižih vrijednosti (prikazano na dijagramu 5.6), kako je već viđeno kod modela linearne regresije i stroja potpornih vektora. Za razliku od njih, na dijagramu kumulativne distribucije vjerojatnosti vidi se kako na srednjem dijelu, u vrijednostima od 45% do 70%, kumulativna distribucija predviđenih vrijednosti prati distribuciju izmjerenih. Preostali metrički pokazatelji prosječnih su vrijednosti kad se usporede s ostalim razvijenim modelima, naznačujući kako model može biti korišten za predviđanja ionosferskog kašnjenja, ali i kako postoje bolji modeli za odrađivanje predviđanja, barem u slučaju gdje su kao prediktori korištene samo komponente gustoće geomagnetskog polja.

Osim predstavljene četveroslojne mreže razvijeno je i nekoliko modela s većim brojem slojeva, a za usporedbu s već predstavljenim je uzet jedan peteroslojni model u koji je dodan jedan potpuno povezani skriveni sloj s 32 neurona unutar sebe te jedan sedmeroslojni model koji u usporedbi s predstavljenim ima dodatna tri skrivena sloja u kojima su redom 32, 64 i 32 neurona. Usporedba dobivenih vrijednosti korijena srednje kvadratne pogreške i prilagođenog koeficijenta determinacije definiranih modela prikazana je u tablici 6.3 na kojoj je vidljivo kako modeli s većim brojem slojeva imaju minimalno bolje vrijednosti navedenih kriterija od predstavljenog modela neuronske mreže, ali nedovoljno dobre da pariraju s ostalim modelima razvijenim drugim metodama, pa zbog toga nisu uzeti kao reprezentativni primjerak.

Tablica 6.3 Usporedba rezultata razvijenih modela neuronskih mreža

Model	Broj slojeva	Broj epoha	Broj uzoraka za ažuriranje	RMSE $[TECU]$	Prilagođeni R^2 [%]
1.	5	64	16	17,1892	0,4019
2.	7	128	48	16,8943	0,4204
Finalni	4	50	32	17,2805	0,3953

Računski model neuronske mreže unošenje nelinearnosti u model postiže korištenjem aktivacijskih funkcija poput ReLU, što u paru sa sposobnošću prepoznavanja složenih uzoraka i automatskim učenjem reprezentacije značajki iz podataka osigurava dobru generalizaciju i visoku fleksibilnost modela, pružajući svojevrsni sveobuhvatni okvir za različite vrste analitičkih zadataka, od klasifikacije do regresije. Naznačena fleksibilnost ostvarena je po cijenu visoke računalne zahtjevnosti kod složenijih arhitektura s većim brojem parametara i težeg tumačenja modela. Kod slučajeva gdje je broj slojeva i neurona znatno veći u odnosu na količinu predanih podataka može doći do pretjerane prilagodbe podataka, stoga model u pravilu daje bolje rezultate kada radi na velikim skupovima podataka. Podešavanje svih parametara vremenski je zahtjevan proces, a s obzirom na činjenicu kako svaki pojedini parametar ne mora imati izraziti učinak na konačno predviđanje, pronalazak potrebnih parametara i njihovo namještanje kako bi se pokušala postići čim bolja predviđanja mogu predstavljati problem.

6.7 Usporedba

Na temelju dobivenih vrijednosti srednje apsolutne pogreške i korijena srednje kvadratne pogreške zaključuje se kako model slučajne šume ima najveću otpornost na netipične vrijednosti podataka, ima najmanja odstupanja od izmjerenih vrijednosti te najbolje generalizira podatke, što ga čini najstabilnijim modelom od razvijenih.

Vrijednosti maksimalnog reziduala kod svih modela su značajno visoke, a radi se o predviđanjima kad se modeli susretnu s netipično visokim izmjerenim vrijednostima. Svi modeli ovdje rade pogrešku u procjeni većoj od 280 TEC-a, osim modela gradijentnog pojačavanja kod kojeg je ta greška za nekoliko desetina TEC-a manja. Sukladno tome, utvrđuje se kako model gradijentnog pojačavanja bolje ograničava najveće pogreške u predviđanjima, što ga čini najpouzdanijim modelom pri radu s ekstremnim vrijednostima. Razlog tome je činjenica da model gradijentnog pojačavanja ima i najveći prilagođeni koeficijent determinacije, koji mu omogućuje preciznija predviđanja u različitim uvjetima, zahvaljujući najbolje objašnjenoj varijabilnosti stvarnih vrijednosti TEC-a, odnosno sposobnosti hvatanja složenih obrazaca u opažanjima.

Vrijeme razvoja je najrazličitija dobivena metrika među razvijenim modelima. Izrazito kratko vrijeme treniranja za osnovne modele kao što su linearna regresija i stablo odluke dolazi na račun lošijih vrijednosti preostalih metričkih pokazatelja, dok s druge strane, u slučaju izrazito dugog vremena treniranja kod modela stroja potpornih vektor, utrošeno vrijeme nije opravdano boljim vrijednostima ostalih metričkih pokazatelja.

Prema opisanim metrikama i odrađenoj statističkoj analizi jasno je kako proces izračuna ionosferskog kašnjenja nije jednostavan, već je riječ o kompleksnom postupku na koji utječe veći broj čimbenika poput geomagnetskih uvjeta, solarnih aktivnosti, geografske lokacije i vremena, čiji su odnosi nelinearni, tj. složeni. Navedeno razjašnjava potrebu prelaska sa stacionarnih modela poput Klobucharevog na modele razvijene metodama strojnog učenja koji se mogu bolje prilagođavati stvarnim vrijednostima radi preciznijih i pouzdanijih predviđanja.

Razvijeni modeli potvrda su teze kako je moguće razviti prognostičke modele ionosferskog kašnjenja tehnikama strojnog učenja, koristeći se lokalnim opažanjima komponenti gustoća geomagnetskog polja i vrijednostima TEC-a za treniranje i pro-

vjeru modela. Dobiveni modeli sposobni su predviđati vrijednosti TEC-a s relativno dobrom uspješnošću u prosjeku. Osnovni modeli poput linearne regresije i stabla odluka korisni su za razumijevanje temeljnih odnosa između podataka, no nedostatni su za davanje preciznih i točnih vrijednosti. S druge strane, naprednije metode strojnog učenja uspijevaju uhvatiti složenije obrasce u podatcima čime znatno poboljšavaju izvedbu predviđanja, a među njima posebno se ističu dva najbolja modela; gradijentno pojačavanje i slučajna šuma.

Oba spomenuta modela daju relativno dobra predviđanja za vrijednosti do 100 TEC-a. Drugim riječima, modeli najbolje rezultate daju za vrijeme nisko poremećenih do blago poremećenih uvjeta u ionosferi. Pogrešno predviđene vrijednosti koje su često rezultat procjene modela za netipično visoke vrijednosti izmjerenog TEC-a proizašle iz jako poremećenih uvjeta u ionosferi za posljedicu mogu imati unošenje neispravnih skrivenih težina ili varijabli u model koje negativno utječu na njegovo učenje i razvoj.

Vrijeme potrebno za razvoj modela gradijentnog pojačavanja i slučajne šume moglo bi predstavljati problem u primjeni tih modela izravno na GNSS prijemniku jer je predugačko za kontinuirana ažuriranja kroz kraće vremenske periode. Za razvoj navedenih modela potrebna je i znatna količina računalnih resursa koja možda ne bi bila dostupna u slučaju opisane izvedbe te bi se za razvoj trebala razmotriti neka od alternativnih metoda poput računarstva u oblaku. No korištenje infrastrukture za razvoj u oblaku koje ovisi o internetskoj vezi i gdje se naplaćuje vrijeme korištenja resursa stvaralo bi nove probleme te bi najvjerojatnije rješenje bio pronalazak kompromisa između brzine razvoja i točnosti modela.

6.8 Poboljšanja

Za poboljšanje rada modela slučajne šume trebalo bi detaljnije ispitati koliko promjena neobaveznih parametara korištenih pri izradi modela [32], poput ispravka pomaka ili broja varijabli nasumično odabranih kao kandidata pri svakom dijeljenju, zaista utječu u slučaju predviđanja ionosferskog kašnjenja. Isto je primjenjivo i za model gradijentnog pojačavanja, no kod njega bi uz navedeno trebalo uvesti i mehanizme ograničenja koji bi sprječavali pojavu predviđanja negativnih vrijednosti TEC-a. Kao opcije poboljšanja procjena oba modela mogle bi se još razmotriti i mogućnosti uvođenja dodatnih prediktora, paralelni razvoj modela na više računala/računalnih jezgri za smanjenje vremena potrebnog za njihovo treniranje te korištenje naprednijih varijanti korištenih algoritama, poput *Extreme gradient boosting* (*xgboost*) [37] algoritma u slučaju gradijentnog pojačavanja.

Kako je raspršenost predviđenih točaka, vidljiva kod svih predviđeno-izmjerenih dijagrama modela na slikama 5.1 do 5.6, relativno velika, još jedan od načina poboljšanja procjena, u ovom slučaju svih modela, bila bi dodatna analiza i transformacija opažanja prije razvoja modela. Detaljnijom analizom opažanja utvrdile bi se određene veze između svake instance što bi poslužilo za transformaciju, odnosno klasifikaciju opažanja u npr. tri različita skupa podataka, ovisno o ionosferskim uvjetima u kojima su izmjereni; mirni ionosferski uvjeti, blago poremećeni ionosferski uvjeti i jako poremećeni ionosferski uvjeti. Modeli razvijeni koristeći podatke dodatno obrađene na spomenuti način bili bi bolje prilagođeni specifičnim karakteristikama opažanja unutar svakog skupa, čime bi se povećala preciznost predviđanja.

Poglavlje 7

Zaključak

Globalni navigacijski satelitski sustavi (GNSS) odgovorni su za globalno određivanja položaja, navigacije i mjerenja vremena, a zbog svoje široke primjene u mnogim područjima moderan život bez usluga koje pružaju bio bi nezamisliv. Spomenute usluge do krajnjih korisnika stižu uz pomoć signala odašiljanih preko satelita u svemiru do odgovarajućih prijemnika koji ih potom obrađuju. Na tom putu signali prolaze kroz ionosferu u kojoj se zbog velike količine električki nabijenih čestica događa degradacija signala koja za posljedicu ima kašnjenje satelitskog signala, a samim time i greške u određivanju položaja. Navedeni fenomen naziva se ionosfersko kašnjenje i jedan je od glavnih uzročnika pogreške satelitskog određivanja položaja, a mjeren je kao ukupni sadržaj elektrona, TEC. Pretpostavljena je mogućnost izrade korektivnog modela vrijednosti TEC-a tehnikama strojnog učenja koji bi mogao biti iskorišten za ublažavanje učinka ionosferskog kašnjenja i samim time doveo do poboljšanja usluga zasnovanih na satelitskoj navigaciji.

Skup podataka korišten za treniranje i testiranje modela preuzet je s internetskog repozitorija Figshare [21] i dodatno uređen prije obuke modela. Općenito je za modele definirano da mogu predvidjeti vrijednosti TEC-a korištenjem izmjerenih komponenti gustoće geomagnetskog polja Bx, By i Bz iz prilagođenog skupa poda-

taka kao prediktora. Određeno je nekoliko kriterija za vrednovanje modela, od kojih su glavni predviđeno-izmjereni dijagram, korijen srednje kvadratne pogreške te prilagođeni koeficijent determinacije. Metodama strojnog učenja razvijeno je, objašnjeno i vrednovano šest prognostičkih modela ionosferskog kašnjenja.

Razvijeni modeli su redom model linearne regresije, stabla odluke, gradijentnog pojačavanja, slučajne šume, stroja potpornih vektora i neuronske mreže. Od svih navedenih, najboljima su se pokazali model gradijentnog pojačavanja i model slučajne šume zbog relativno dobrih rezultata kroz sve metrike i vidljivo najtočnijih predviđanja vrijednosti TEC-a. Za oba spomenuta modela predloženo je nekoliko poboljšanja koja bi mogla utjecati na njihovu izvedbu i poboljšati preciznost predviđanja, kao što su klasifikacija ulaznih podataka te bolja definicija neobaveznih parametara predanih funkciji za izradu modela i detaljnija provjera njihovog utjecaja na krajnje rezultate.

Moguća daljnja istraživanja uključuju razvijanje novih prognostičkih modela ionosferskog kašnjenja korištenjem nekih drugih metoda strojnog učenja, poput Bayesove regresije [38] ili naprednijih varijanti algoritama modela slučajne šume i gradijentnog pojačavanja, te usporedbu istih sa standardnim (Klobucharevim) modelom, kao i rezultatima modela priloženih unutar diplomskog rada.

Literatura

[1] *ESA Satellite Navigation*, s Interneta, Pristupljeno 08.08.2024. adresa: https://www.esa.int/Applications/Satellite_navigation/Who_benefits_some_practical_applications.

[2] *Global Positioning System*, s Interneta, Pristupljeno 08.08.2024. adresa: https://www.gps.gov/.

[3] *GSC Europe Galileo*, s Interneta, Pristupljeno 08.08.2024. adresa: https://www.gsc-europa.eu/galileo/what-is-galileo.

[4] *GLONASS*, s Interneta, Pristupljeno 08.08.2024. adresa: https://glonass-iac.ru/en/about_glonass/.

[5] *BeiDou*, s Interneta, Pristupljeno 08.08.2024. adresa: http://en.beidou.gov.cn/.

[6] *EUSPA GNSS*, s Interneta, Pristupljeno 08.08.2024. adresa: https://www.euspa.europa.eu/eu-space-programme/galileo/what-gnss.

[7] *Advanced Navigation - GNSS and Satellite Navigation*, s Interneta, Pristupljeno 08.08.2024. adresa: https://www.advancednavigation.com/tech-articles/global-navigation-satellite-system-gnss-and-satellite-navigation-explained/.

[8] J. S. Subirana, J. M. J. Zornoza i M. Hernandez-Pajares, *GNSS DATA PROCESSING - Volume I: Fundamentals and Algorithms*, K. Fletcher, ur. ESA Communications, 2013.

[9] *NeQuick Ionospheric Model*, s Interneta, Pristupljeno 08.08.2024. adresa: https://gssc.esa.int/navipedia/index.php/NeQuick_Ionospheric_Model.

[10] *Klobuchar Ionospheric Model*, s Interneta, Pristupljeno 08.08.2024. adresa: `https://gssc.esa.int/navipedia/index.php?title=Klobuchar_Ionospheric_Model`.

[11] K. Davies, *Ionospheric Radio* (IEE electromagnetic waves series). London: IET, 1990.

[12] W. P. Bradford i J. J. Spilker, *Global positioning system: Theory and Application, Theory and applications*. Washington, DC: American Inst. of Aeronautics i Astronautics, 1996.

[13] J. A. Klobuchar, „Ionospheric Time-Delay Algorithm for Single-Frequency GPS Users," *IEEE Transactions on Aerospace and Electronic Systems*, 1987.

[14] *IBM - what is ML?* s Interneta, pristupljeno 08.08.2024. adresa: `https://www.ibm.com/topics/machine-learning`.

[15] E. Alpaydın, *Introduction to machine learning*. Cambridge, Mass. [u.a.]: MIT Press, 2004.

[16] A. Nigusie, A. Tebabal i R. Galas, „Modeling Ionospheric TEC Using Gradient Boosting Based and Stacking Machine Learning Techniques," *Space Weather*, sv. 22, br. 3, 2024. DOI: `10.1029/2023SW003821`.

[17] R. Filjar i dr., „Predictive Model of Total Electron Content during Moderately Disturbed Geomagnetic Conditions for GNSS Positioning Performance Improvement," 2020. DOI: `10.23919/FUSION45008.2020.9190264`.

[18] J. Tang i dr., „An Approach for Predicting Global Ionospheric TEC Using Machine Learning," *Remote Sensing*, sv. 14, br. 7, str. 1585, 2022. DOI: `10.3390/rs14071585`.

[19] *Posit RStudio*, s Interneta, Pristupljeno 08.08.2024. adresa: `https://posit.co/products/open-source/rstudio/`.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. adresa: `https://www.R-project.org/`.

[21] I. Heđi i dr., „Hedji, Cirikovic, Borkovic, Filjar, JCIEES 2023 manuscript, Supplementary material," 2023. DOI: `10.6084/m9.figshare.22579786.v1`.

[22] I. Heđi i dr., „A method for Assemblage of an Open Access Data Set for Research in Geomagnetic Effects on GPS/GNSS Ionospheric Delay in Sub-equatorial Regions," *The Journal of CIEES*, sv. 3, br. 1, str. 7–11, 2023. DOI: 10.48149/jciees.2023.3.1.1.

[23] *NASA Earth Data*, s Interneta, Pristupljeno 08.08.2024. adresa: https://www.earthdata.nasa.gov/.

[24] G. K. Seemala, „Estimation of ionospheric total electron content (TEC) from GNSS observations," *Atmospheric Remote Sensing*. Elsevier, 2023., str. 63–84. DOI: 10.1016/b978-0-323-99262-6.00022-5.

[25] *INTERMAGNET repositories*, s Interneta, Pristupljeno 08.08.2024. adresa: https://github.com/orgs/INTERMAGNET/repositories.

[26] J. M. Chambers, ur., *Statistical models in S*. Pacific Grove, Calif.: Wadsworth i Brooks, 1992.

[27] L. Breiman i dr., *Classification and regression trees*. Boca Raton, Fla. [u.a.]: Chapman i Hall/CRC, 1998.

[28] T. Therneau i B. Atkinson, *rpart: Recursive Partitioning and Regression Trees*, R package version 4.1.23, prosinac 2013. DOI: 10.32614/cran.package.rpart. adresa: https://cran.r-project.org/web/packages/rpart/.

[29] J. H. Friedman, „Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, br. 5, 2001. DOI: 10.1214/aos/1013203451.

[30] G. Ridgeway i G. Developers, *gbm: Generalized Boosted Regression Models*, R package version 2.2.2, lipanj 2024. DOI: 10.32614/cran.package.gbm. adresa: https://cran.r-project.org/web/packages/gbm/.

[31] L. Breiman, „Random Forests," *Machine Learning*, br. 1, str. 5–32, 2001. DOI: 10.1023/A:1010950718922.

[32] L. Breiman i dr., *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*, R package version 2.2.2, svibanj 2022. DOI: 10.32614/cran.package.randomforest. adresa: https://cran.r-project.org/web/packages/randomForest/.

[33] C.-C. Chang i C.-J. Lin, „LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, br. 3, str. 1–27, 2011. DOI: 10.1145/1961189.1961199.

[34] D. Meyer i dr., *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, R package version 1.7-14, prosinac 2023. DOI: 10.32614/cran.package.e1071. adresa: https://cran.r-project.org/web/packages/e1071/.

[35] J. Allaire i F. Chollet, *keras: R Interface to "Keras"*, R package version 2.15.0, travanj 2024. DOI: 10.32614/cran.package.keras. adresa: https://cran.r-project.org/web/packages/keras/.

[36] J. Allaire i Y. Tang, *tensorflow: R Interface to "TensorFlow"*, R package version 2.16.0, travanj 2024. DOI: 10.32614/cran.package.tensorflow. adresa: https://cran.r-project.org/web/packages/tensorflow/.

[37] T. Chen i dr., *xgboost: Extreme Gradient Boosting*, R package version 1.7.8.1, srpanj 2024. DOI: 10.32614/cran.package.xgboost. adresa: https://cran.r-project.org/web/packages/xgboost/.

[38] P.-C. Bürkner, *brms: Bayesian Regression Models using "Stan"*, R package version 2.21.0, ožujak 2024. DOI: 10.32614/cran.package.brms. adresa: https://cran.r-project.org/web/packages/brms/.

[39] M. Petranović, *R code for models.zip*, 2024. DOI: 10.6084/M9.FIGSHARE.26963689.V1.

Pojmovnik

AI Artificial Intelligence. 6

GBM Gradient Boosting Machine. 21

GNSS Global Navigation Satellite System. 1, 4–6, 8–10, 12, 54, 56

GPS Global Positioning System. 1, 2, 5, 12, 34

IDE Integrated Development Environment. 11

IGS International GNSS Service. 12

INTERMAGNET International Real-time Magnetic Observatory Network. 12

MAE Mean Absolute Error. 29, 33, 34, 43

ML Machine Learning. 6

MSE Mean Square Error. 29

PNT Positioning, Navigation and Timing. 1

PPP Precise Point Positioning. 2

RMSE Root Mean Square Error. 6, 32–34, 43

RTK Real-time Kinematics. 2

SVM Support Vector Machine. 7

SVR Support Vector Regression. 25

TEC Total Electron Content. 2, 4, 5, 7–9, 12–15, 18, 20, 29, 32, 43–46, 48, 50, 51, 53–57

Sažetak

Ionosfersko kašnjenje jedan je od najvećih utjecajnih čimbenika na kašnjenje satelitskih signala globalnih navigacijskih satelitskih sustava (GNSS), a mjeri se brojem ukupnog sadržaja elektrona (TEC). Unutar diplomskog rada pretpostavljena je mogućnost izrade prognostičkih modela GPS ionosferskog kašnjenja s komponentama geomagnetskog polja kao prediktorima, metodama strojnog učenja, s ciljem smanjivanja prvotno spomenutog utjecaja na satelitske signale. Ukratko je opisan fenomen ionosferskog kašnjenja, zajedno s primjerima dosadašnjeg istraživanja vezanim uz razvoj korektivnih modela vrijednosti TEC-a, poput stacionarnog Klobucharevog modela te pristup izrade prilagodljivom modelu tehnikama strojnog učenja. Definirana je metodologija rada, od dobivanja, analize i dodatne obrade korištenih ulaznih podataka do postavljanja kriterija na temelju kojih su razvijeni modeli vrednovani. Detaljno je razrađen postupak razvoja modela, navedene su prednosti i mane svakog modela u danom kontekstu te su na kraju međusobno uspoređeni. Rezultat diplomskog rada šest je razvijenih modela namijenjenih predviđanju vrijednosti TEC-a: model linearne regresije, stabla odluke, gradijentnog pojačavanja, slučajne šume, stroja potpornih vektora i neuronskih mreža.

Ključne riječi **— satelitska navigacija, ionosfersko kašnjenje, strojno učenje, ukupni sadržaj elektrona(TEC)**

Abstract

Ionospheric delay, measured by the Total Electron Content (TEC) number, is one of the largest influencing factors for the delay of satellite signals used by Global Navigation Satellite Systems (GNSS). Within this master's thesis, the possibility of developing a machine learning-based predictive model of GPS ionospheric delay with geomagnetic field components as predictors is hypothesized, which aims to reduce the aforementioned influence on satellite signals. The phenomenon of ionospheric delay is briefly described, together with examples of previous research related to the development of corrective models for the TEC value, such as the stationary Klobuchar model and the approach to creating an adaptive model using machine learning techniques. Work methodology is defined from the acquisition, analysis and

additional processing of the used input data to the setting of criteria based on which the developed models are evaluated. The model development procedure is detailed, with the advantages and methods of each model in a given context listed, and at the end, models are compared with each other. The result of the thesis is six developed models intended for predicting the value of TEC: linear regression, decision tree, gradient boosting, random forest, support vector machine and neural network.

Keywords **— satellite navigation, ionospheric delay, machine learning, total electron content (TEC)**

Dodatak A

Programska podrška

Pristup računalnom kodu za izradu prognostičkih modela ionosferskog kašnjenja razvijenim u programskom okruženju za statističko računarstvo R koji je korišten u diplomskom radu omogućen je na internetskom repozitoriju otvorenog pristupa Figshare preko sljedeće poveznice:

https://doi.org/10.6084/m9.figshare.26963689.v1 [39]

Podatci potrebni za razvoj modela iz gore priložene programske podrške dostupni su na poveznici:

https://doi.org/10.6084/m9.figshare.22579786.v1 [21]