Potraga za egzoplanetima primjenom algoritama u programskom jeziku Python

Pajas, Matija

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

Matija Pajas

POTRAGA ZA EGZOPLANETIMA PRIMJENOM ALGORITAMA U PROGRAMSKOM JEZIKU PYTHON

Diplomski rad

Zagreb, 2021.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

INTEGRIRANI PREDDIPLOMSKI I DIPLOMSKI SVEUČILIŠNI STUDIJ
FIZIKA I INFORMATIKA; SMJER NASTAVNIČKI

Matija Pajas

Diplomski rad

Potraga za egzoplanetima primjenom algoritama u programskom jeziku Python

Voditelj diplomskog rada: izv. prof. dr. sc. Goranka Bilalbegović

Ocjena diplomskog rada:_____

Povjerenstvo: 1._____

2._____

3._____

Datum polaganja: _____

Zagreb, 2021.

Zahvaljujem svojoj mentorici Goranki Bilalbegović na pruženom znanju, strpljenju, idejama i velikoj pomoći pri pisanju ovog diplomskog rada.
Hvala prijateljima i kolegama koji su mi bili podrška i uljepšali mi razdoblje studiranja.
Najveću zahvalu dugujem svojoj obitelji, a posebno zahvaljujem bratu i roditeljima bez kojih ništa od ovog ne bi bilo moguće. Hvala vam na strpljenju, što ste uvijek bili uz mene i pružili mi veliku podršku.

Sažetak

Egzoplaneti su planeti koji se nalaze izvan Sunčevog sustava. Oni su fascinantni astronomima i amaterima koji žele pronaći planete s izvanzemaljskim oblicima života. Snimanjem zvijezda pomoću satelitskih teleskopa, kao što su Kepler i TESS, pronašlo se puno egzoplaneta. Ručno analiziranje teleskopskih podataka vrlo je težak i dugotrajan posao. Moguće je smanjiti broj kandidata za ručnu provjeru primjenom strojnog učenja i drugih naprednih računalnih metoda. Koristili smo logističku regresiju i stabla odlučivanja za klasifikaciju zvijezda na one koje imaju i one koje nemaju egzoplanete. Usporedili smo rezultate te dvije metode strojnog učenja. U metodičkom dijelu raspravljamo važnost astronomskih projekata u znanosti za građanstvo i njihovu moguću ulogu u školama. Pored toga predstavljena je nastavna priprema za održavanje sata o Newtonovom zakonu gravitacije.

Ključne riječi: egzoplaneti, strojno učenje, nadzirano strojno učenje, logistička regresija, stabla odlučivanja, znanost za građanstvo, nastava fizike za srednje škole: Newtonov zakon gravitacije

Search for exoplanets using algorithms in programming language Python

Abstract

Exoplanets are planets outside the Solar system. They are fascinating for astronomers and amateurs wanting to find planets with signs of extraterrestrial life forms. By capturing starlight using space telescopes, such as Kepler and TESS, many exoplanets have been discovered. Manual analysis of telescope data is a very tedious and long-lasting job. It is possible to reduce data for manual analysis using machine learning and other advanced computational methods. We used logistic regression and decision trees to classify stars by the presence of exoplanets and compared the results of both algorithms. In the methodical section we discuss the importance of astronomy projects in citizen science, as well as their possible role in schools. We also present the preparation material for teaching Newton's law of universal gravitation.

Keywords: exoplanets, machine learning, supervised machine learning, logistic regression, decision trees, citizen science, high school physics teaching: Newton's law of gravitation

Sadržaj

1 Uvod

Prvi egzoplaneti otkriveni su 1990. godine, a danas se možemo pohvaliti s više od 4000 potvrđenih i još barem dvostruko više kandidata [1]. Ovo područje istraživanja doživjelo je svoje prve velike uspjehe lansiranjem satelitskog teleskopa Kepler [2]. Zadnjih godina potraga za egzoplanetima postaje sve popularnija i povezuje mnoge amatere i znanstvenike koji žele pronaći nove planete sa znakovima života. Istraživanja egzoplaneta daju otvoren pristup svim prikupljenim podacima te se pojavio velik broj samostalnih analiza podataka raznim naprednim računalnim metodama uključujući i strojno učenje. Ovaj diplomski rad bavi se klasifikacijom podataka o egzoplanetima metodama strojnog učenja.

Polazni korak je urađen po uzoru na jednu takvu analizu [3] koja koristi podatke objavljene na Kaggle stranici, najvećoj arhivi podataka specijaliziranih za upotrebu metoda podatkovne znanosti. Automatizirana klasifikacija primjenom strojnog učenja olakšava svakodnevni posao astronoma generiranjem mogućih kandidata iz vrlo velikih baza podataka. Teleskop Kepler je na primjer ukupno snimio oko pola milijuna različith zvijezda [2]. U radu predlažemo drugačiji pristup obradi podataka nakon koje su korištena dva standardna algoritma za klasifikaciju s najboljim rezultatima: logistička regresija i stablo odlučivanja. Pomoću njih uspjeli smo točno pronaći skoro sve egzoplanete u skupu podataka za testiranje.

U drugom poglavlju opisani su egzoplaneti i motivacija za njihovom potragom te metode detekcije. U trećem poglavlju opisujemo nadzirano strojno učenje i algoritme koje koristimo u radu. U četvrtom poglavlju su opisani korišteni alati i postupak obrade podataka te su predstavljeni rezultati metoda strojnog učenja. Nakon Zaključka slijedi metodičko poglavlje koje je posvećeno projektu Planet Hunters Tess, jednom od najpopularnijih projekata znanosti za građanstvo iz područja astronomije. U tom poglavlju opisujemo značaj takvih projekata u popularizaciji znanosti i predlažemo njihovu primjenu u školi za poticanje interesa i razvoj interdisciplinarnih vještina učenika. U metodičkom poglavlju također je predstavljena nastavna priprema za izvođenje sata o Newtonovom zakonu gravitacije.

2 Egzoplaneti

2.1 O egzoplanetima

Egzoplaneti ili ekstrasolarni planeti (engl. *exoplanet* ili *extrasolar planet*) su planeti koji se nalaze izvan Sunčevog sustava. Danas (14. 6. 2021.) ih je potvrđeno 4401 i većina se nalazi u orbiti zvijezda u manjem području Mliječne staze blizu Sunčevog sustava [1]. Uz njih postoji još 6625 kandidata koje je potrebno detaljnije provjeriti s nekoliko metoda kako bi ih se potvrdilo, ili odbacilo. Potraga za egzoplanetima brzo napreduje te se broj potvrđenih planeta i kandidata svakodnevno mijenja. Prvi egzoplaneti otkriveni su 1990., a prije 10 godina bilo ih je samo oko pet stotina.

Broj otkrivenih egzoplaneta je naglo porastao nakon što je 2009. godine u Zemljinu orbitu lansiran svemirski teleskop Kepler [2]. Tako je započelo moderno doba potrage za egzoplanetima. Postavljanjem teleskopa u orbitu iznad Zemlje omogućeno je konstantno snimanje područja svemira na duže vrijeme. Teleskopi na površini Zemlje su ograničeni njenom rotacijom oko svoje osi. Tijekom 9 godina i 7 mjeseci rada, Kepler je snimio više od pola milijuna zvijezda i pronašao 2662 egzoplaneta. Nasljednik Keplera s novom misijom snimanja šireg područja svemira je teleskop TESS (Transiting Exoplanet Survey Satellite) lansiran 2018. godine koji, kao i Kepler, traži planete tranzitnom metodom [4].

Motivacija znanstvenika ne proizlazi samo iz želje za katalogom planeta koji bi jednog dana čovječanstvu svakako mogli biti od koristi barem kao izvor resursa. Ljudi žele odgovoriti na staro, egzistencijalno pitanje: ”Jesmo li jedini život u svemiru?”. Odgovor nije jednostavan jer ne znamo kakvi oblici života mogu postojati izvan planeta Zemlje. Zbog toga su znanstvenicima najzanimljiviji planeti slični Zemlji u takozvanoj naseljivoj zoni koja je pogodna za život koji poznajemo. Postoji nekoliko metoda kojima se za neki planet određuje koliko je sličan Zemlji. Prolaskom svjetlosti sa zvijezde kroz atmosferu planeta moguće je analizom spektra odrediti sastav unutrašnjosti planeta te njegove atmosfere i pronaći plinove koji postoje na Zemlji, kao što su to kisik, dušik, metan i ugljikov dioksid [1]. Pomoću udaljenosti planeta od zvijezde i njegove mase, uz poznavanje sastava atmosfere, također je moguće procijeniti prosječnu temperaturu i mogućnost postojanja vode u tri agregatna stanja. To su neki od preduvjeta za život kakav poznajemo na Zemlji. Nije sigurno da se na planetu s odgovarajućim osobinama razvio život, već samo da bi vjerojatno mogao. A ako se

i razvio život, to ne znači da je taj oblik života razvijen tako da može komunicirati s ljudima. Sa sve većim brojem pronađenih egzoplaneta bliže smo pronalasku života, a time i boljoj procjeni koliko je vjerojatno da se život razvije na njima. Potraga za egzoplanetima je novo i uzbudljivo područje koje se počelo brzo razvijati tek zadnjih tridesetak godina, a još nismo provjerili ni kap nezamislivo prostranog svemira. U otvorenom projektu Planet Hunters TESS [5], kome svi mogu pristupiti, potraga za egzoplanetima povezuje velik broj znanstvenika i amatera. Što nas još čeka na udaljenim planetima možemo samo nagađati.

2.2 Metode detekcije egzoplaneta

Postoji nekoliko metoda kojima se danas pronalaze egzoplaneti [6]. Najjednostavnija od njih je direktno slikanje sa Zemlje. Ovom metodom pronađeno je vrlo malo planeta jer je reflektirana svjetlost udaljenih planeta prigušena jakom svjetlosti njihovih zvijezda. Smetnje se mogu smanjiti korištenjem koronografa (engl. coronograph), ili specijalnih sjenila (engl. starshade) koji blokiraju dio svjetlosti, te digitalnom obradom slika.

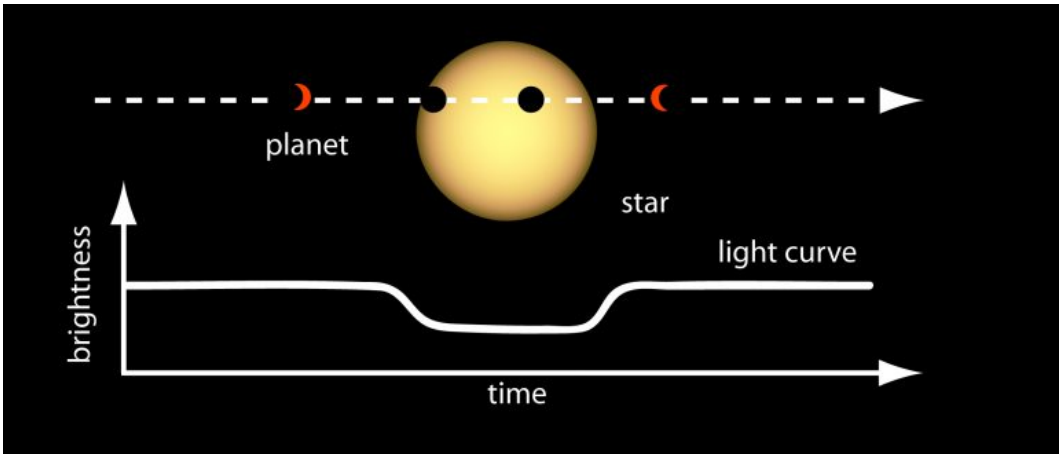
Nešto više planeta pronađeno je pomoću gravitacijskih leća. Planet i njegova zvijezda se mogu gibati između udaljene zvijezde i Zemlje te svojim gravitacijskim poljem savijati putanju svjetlosti daleke zvijezde. Djeluju kao gravitacijska leća fokusirajući svjetlost u jednu točku. Dolazi do privremenog povećanja intenziteta svjetlosti koja sa udaljene zvijezde stiže u blizinu Zemlje i tako se otkrivaju egzoplaneti.

Dvije najuspješnije metode, metoda mjerenja radijalne brzine i detekcija prolaska planeta ispred zvijezde (tzv. tranzitna metoda), zajedno su zaslužne za otkriće 95.1% ukupnog broja dosad potvrđenih egzoplaneta (14. 6. 2021.).

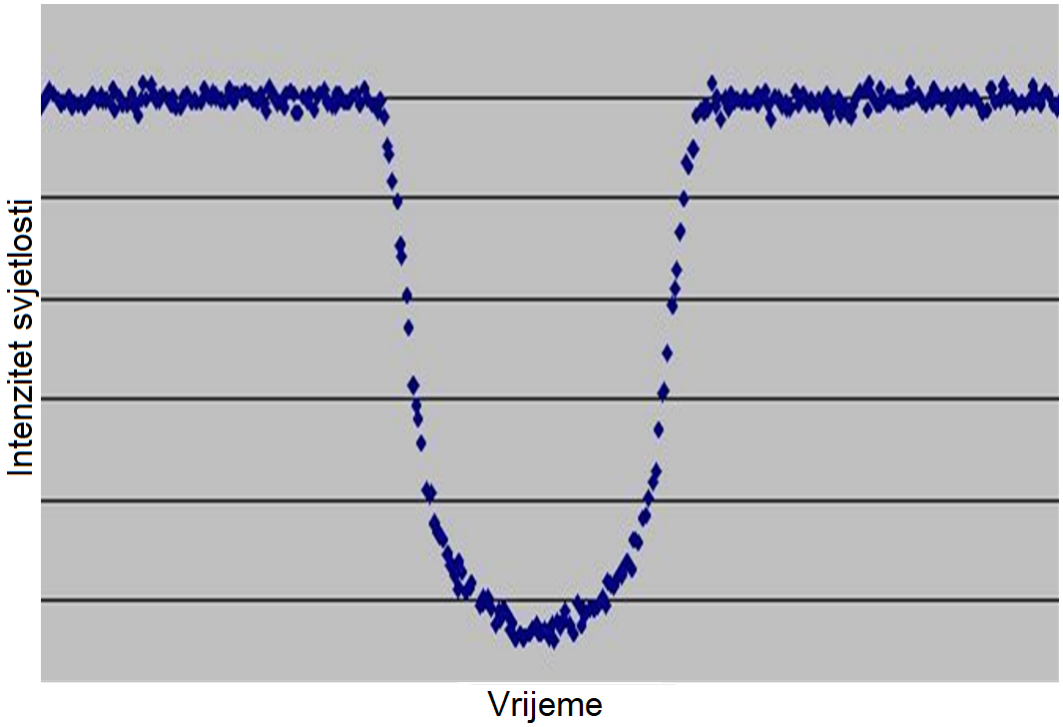
Metoda radijalne brzine se zasniva na promjenama u gibanju zvijezde koje su izazvane prisustvom planeta. Takvo gibanje vodi do promjene boja u spektru zvijezde. Planeti gravitacijski utječu na zvijezdu. Zvijezde zbog toga orbitiraju oko centra mase svog planetarnog sustava. Za jako udaljene zvijezde nije lako primijetiti njihovo gibanje. Moguće je detektirati promjene u valnoj duljini svjetlosti sa zvijezde koje se događaju zbog Dopplerovog efekta. Kada se zvijezda tijekom svog gibanja izazvanog prisustvom planeta udaljava od Zemlje valna duljina svjetlosti koju snimamo se povećava i tu pojavu zovemo crveni pomak (engl. redshift). U suprotom se valna

duljina svjetlosti smanjuje, odnosno događa se plavi pomak (engl. blueshift).

Najuspješnija metoda sa 75.8% ukupnog broja potvrđenih planeta je tranzitna metoda. Snimanjem zvijezda kroz duže vrijeme moguće je analizirati kako njihov intenzitet svjetlosti ovisi o vremenu. Može se dogoditi da teleskop snimi smanjenje tog intenziteta izazvano prolazom egzoplaneta. Metoda, prikazana je na Slici 2.1, se zove tranzitna jer se planeti mogu detektirati jedino ako prolaze između zvijezde i teleskopa. Jedna takva promjena intenziteta prikazana je na Slici 2.2 i rezultat je prolaza egzoplaneta Kepler-6b između zvijezde Kepler 6 i Zemlje. Tranzitna metoda je ograničena činjenicom da planeti moraju imati orbitu orijentiranu tako da prolaze između zvijezde i teleskopa kako bi se mogao detektirati pad intenziteta svjetlosti. Ograničena je također periodom snimanja, jer smo najsigurniji u prolazak planeta kada se on dogodi više puta. Osim toga, kako bismo detektirali prolaske koji su se dogodili samo jednom tijekom snimanja moramo imati sreće da se prolazak dogodi upravo za vrijeme snimanja jer planet može imati duži period orbite te ga možemo lako propustiti. Metoda ima pristranost prema velikim planetima koje imaju male periode orbite i nalaze se na malim udaljenostima od svojih zvijezda. Takvi planeti zbog svoje veličine i blizine zvijezda blokiraju više svjetlosti, a s manjim periodom udubine u grafu intenziteta svjetlosti su češće [1]. Osim što je tranzitna metoda vrlo uspješna u pronalaženju egzoplaneta, zbog snimanja svjetlosti prilikom prolaska planeta moguće je odmah analizirati sastav njegove atmosfere, odrediti temperaturu, a na temelju dubine i širine udubina u grafu mogu se procijeniti dimenzije planeta.



Slika 2.1: Prikaz tranzitne metode detekcije egzoplaneta [7].



Slika 2.2: Promjena intenziteta svjetlosti zvijezde Kepler-6 koju je proizveo egzoplanet Kepler-6b prolazom između zvijezde i teleskopa Kepler [8].

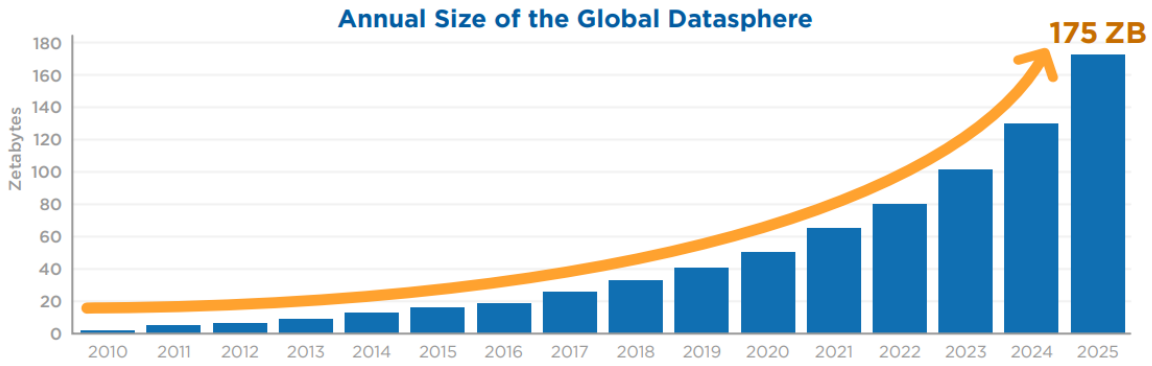
3 Osnove strojnog učenja

3.1 O strojnom učenju

Strojno učenje je grana umjetne inteligencije čiji je zadatak proučavanje i izrada algoritama koji mogu učiti iz podataka. Nešto precizniju definiciju dao je Tom Mitchell 1997. godine: ”Program uči iz iskustva (I) ako, nakon što uzme u obzir mjeru preciznosti (P) za neki zadatak (Z), njegova preciznost na Z, nakon mjerenja P, se poveća s iskustvom I.” [9]. Ideja o automatskom prilagođavanju programa na ulazne podatke nije nova. Međutim, do vrlo brzog razvoja strojnog i dubokog strojnog učenja (engl. deep machine learning) [10] je došlo u zadnjih dvadesetak godina s razvojem snažnih računala. Raste zainteresiranost za primjene strojnog učenja u znanosti, gospodarstvu i svakodnevnom životu. Danas smo strojnom učenju izloženi svakodnevno na internetu ili u primjeni mobilnih aplikacija. Koristi ga YouTube kako bi nam preporučio najbolje filmove, Facebook najbolje reklame, LinkedIn poslove koji će nas najvjerojatnije zanimati. Veliki dućani koriste strojno učenje u analizi proizvoda koje je najbolje staviti na zajedničku policu na temelju navika kupaca kako bi prodali što više proizvoda.

Prednost koju strojno učenje ima nad klasičnim programiranjem je to što programi ne moraju biti eksplicitno programirani kako bi izvršavali svoje zadatke. Umjesto ručnog pisanja uvjetnog grananja za svaku moguću kombinaciju, što može biti jako komplicirano, programer može koristiti strojno učenje i naučiti algoritam da prepoznaje obrasce u kompliciranim zadacima. Rješenje možda neće biti savršeno, ali će biti jako korisno kod podataka u kojima čovjek ne može naći određeno pravilo. Tako je na primjer s pojavom neuronskih mreža postalo moguće prepoznavati lica ili glasove. Ne postoji posebno pravilo koje je moguće programirati tako da program prepozna glas ili nečije lice. Čak i kada bi to bilo moguće, trebalo bi prepoznati karakteristike svakog pojedinog glasa i lica. Problem je kad treba prepoznati stotine, tisuće, ili milijune uzoraka. Očito je da želimo neko generalnije rješenje koje će se moći primijeniti na široku klasu problema. Neuronske mreže će moći prepoznati obrasce ako ih se uči na dovoljno velikom skupu podataka. Problem u primjeni strojnog učenja je činjenica da je potrebno puno podataka za njihovu učinkovitost. Osim toga gotovo uvijek je potrebno podatke prilagoditi i obraditi prije upotrebe algoritama, ali ne uvijek na očit način. Međutim, računala su u stanju raditi s velikom količinom podataka puno

brže nego ljudi. Količina podataka danas raste brže nego ikad u povijesti. Američka tvrtka Seagate, koja se bavi rješenjima za masovno skladištenje podataka, provela je istraživanje u kojem predviđa da će količina podataka na globalnoj razini porasti sa 33 zetabajta u 2018. godini na 175 zetabajta do 2025. godine (vidi Sliku 3.1) [11].



Slika 3.1: Godišnja količina podataka na globalnoj razini [11].

Osim potrebe za velikom količinom podataka, također je važno prepoznati kada je strojno učenje najbolji izbor za problem koji rješavamo. Strojno učenje se može podijeliti na četiri glavne kategorije [10]:

- Nadzirano strojno učenje (engl. *supervised learning*) – Naziv je dobilo po tome što čovjek nadzire algoritam tako što unaprijed pripremi željene izlaze za pojedine ulazne varijable. Na taj način algoritam uči. Glavna zadaća nadziranog učenja je što bolja generaliziracija postupka na nove podatke. Neki tipični problemi koje rješava ova vrsta algoritama su regresija i klasifikacija.

- Nenadzirano strojno učenje (engl. *unsupervised learning*) – Suprotno nazdiranom učenju, algoritmi nenadziranog strojnog učenja nemaju pripremljene željene izlaze već im je zadaća za zadani skup podataka samostalno odrediti zajedničke karakteristike. Neke od bitnih zadaća su im grupiranje (engl. *clustering*), reprezentacija i vizualizacija u manjim dimenzijama te prepoznavanje anomalija, to jest novih podataka koji značajno odstupaju od naučenih podataka.

- Polunadzirano strojno učenje (engl. *semisupervised learning*) – To su algoritmi koji implementiraju kombinaciju nadziranog i nenadziranog strojnog učenja. Ako set podataka nije potpuno označen, na njemu nema smisla koristiti nadzirano učenje. No ako taj set podataka prvo grupira algoritam nenadziranog učenja, moguće je označiti podatke za nadzirani dio algoritma.

- Učenje s podrškom (engl. *reinforcement learning*) – To je vrsta algoritama koji uče uz interakciju s okolinom u realnom vremenu. Algoritam postiže željeno ponašanje na temelju nagrade i kazne koje se računaju odmah nakon postupanja algoritma kao odgovor na okolinu. Cilj algoritma je postići što veći broj nagrada, a što manji broj kazni. Ova vrsta algoritama na primjer može koristiti robotima kako bi naučili hodati, ili autima kako bi samostalno vozili.

U ovom radu se koristi nekoliko algoritama nadziranog strojnog učenja koji su se pokazali pogodni za otkrivanje egzoplaneta u skupu podataka koji sadrži kandidate.

3.2 Nadzirano strojno učenje

Nadzirano strojno učenje, kao što je već spomenuto, uči pomoću podataka koji su unaprijed povezani sa željenim rezultatima. Svaki podatak karakteriziran je svojim značajkama (engl. *features*) i pridružene su mu oznake (engl. *labels*) koje predstavljaju pravilno rješenje za određeni podatak. Zadaća nadziranog algoritma je dati što bolje rješenje na temelju značajki. Skup ulaznih podataka se može zapisati kao vektor $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, pri čemu je n broj ulaznih podataka. Svaki vektor $\mathbf{x}_i$ sadrži numeričke vrijednosti značajki za i-ti podatak. Vektor $\mathbf{y}$ sadrži oznake, odnosno izlazne numeričke vrijednosti, pri čemu i-ti element tog vektora sadrži oznaku i-tog ulaznog podatka. Tako definirani ulazni podaci mogu se zapisati kao matrica $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Značajke su povezane s oznakama funkcijom koja ovisi o skupu parametara $\mathbf{w}$ koji predstavljaju težinu pojedine značajke:

$$\mathbf{y} = f(\mathbf{X}; \mathbf{w}) \tag{3.1}$$

odnosno za pojedini podatak:

$$y_i = f(\mathbf{x}_i) \tag{3.2}$$

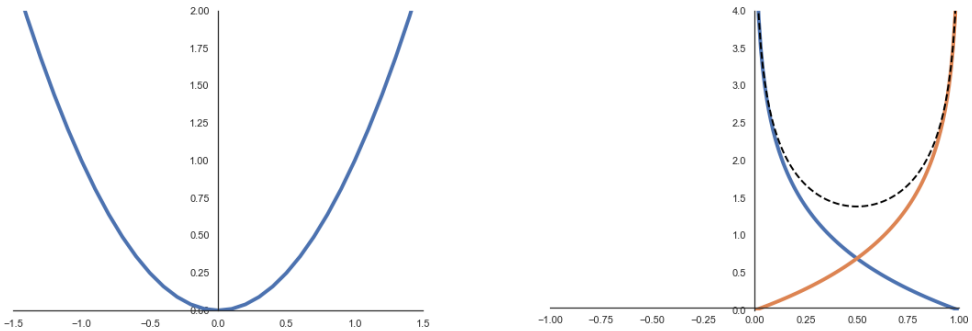
Izabrati funkciju znači odabrati algoritam s funkcijom koja je prikladna za problem

koji rješavamo. Za svaki set značajki funkcija će ispisivati procjenu rezultata, a algoritam će namještati parametre $\mathbf{w}$ nakon izračunate procjene. To je proces zbog kojeg je strojno učenje dobilo svoje ime. Cilj nadziranog strojnog učenja je pronaći što bolju funkciju f koja će predviđati oznake $\mathbf{y}$ za zadane ulazne podatke $\mathbf{x}$. Još bitnije od toga, želimo da ta funkcija što bolje predviđa vrijednosti novih još nepoznatih podataka. U trenutku kada je algoritam naučen zvat ćemo tu funkciju h ili hipoteza, a ona će kao izlaz davati predviđenu vrijednost naučenog modela $\hat{y}_i = h(\mathbf{x}_i)$.

Funkcija troška i gradijentni spust Važno je objasniti kako ovi algoritmi uče i mijenjaju parametre $\mathbf{w}$ u jednažbi 3.1. Za svaki algoritam definira se ciljna funkcija (engl. *objective function*) koja se, ovisno o željenom ishodu, mora minimizirati ili maksimizirati za postizanje optimalnog rješenja. Tako se na primjer u slučaju linearne regresije kao ciljna funkcija koristi srednja kvadratna pogreška (engl. *mean square error*, MSE) [10]:

$$\mathrm{MSE}(\mathbf{X}, h_{\mathbf{w}}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}^T\mathbf{x}_i - y_i\right)^2 \tag{3.3}$$

Cilj je smanjiti pogrešku algoritma te je srednju kvadratnu pogrešku potrebno minimizirati i zbog toga je zovemo funkcija troška (engl. *cost function*). Logistička regresija, koja je algoritam blizak linearnoj regresiji, ima funkciju troška koja se razlikuje od srednje kvadratne pogreške. Međutim, kao što je prikazano na Slici 3.2, obje funkcije troška su konveksne i moguće im je garantirano pronaći minimum.



(a) (b)

Slika 3.2: Kvadratna funkcija (a) i funkcija troška logističke regresije (b).

Jedan od postupaka za izračunavanje minimuma je gradijentni spust (engl. *gradient descent*) [10]. Kao što sam naziv sugerira, za funkciju se izračuna gradijent po

parametrima $\mathbf{w}$. Taj gradijent pokazuje smjer najmanje vrijednosti funkcije, a zatim se napravi korak u smjeru gradijenta. Koraci se ponavljaju dok gradijentni spust ne dođe do minimuma funkcije, a to će biti onda kada vrijednost gradijenta postane zanemarivo mala i bliska nuli. U slučaju srednje kvadratne pogreške gradijent je [10]:

$$\nabla_{\mathbf{w}}\mathrm{MSE}(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_1}\mathrm{MSE}(\mathbf{w}) \\ \frac{\partial}{\partial w_1}\mathrm{MSE}(\mathbf{w}) \\ \vdots \\ \frac{\partial}{\partial w_n}\mathrm{MSE}(\mathbf{w}) \end{bmatrix} \tag{3.4}$$

pri čemu je pojedina parcijalna derivacija:

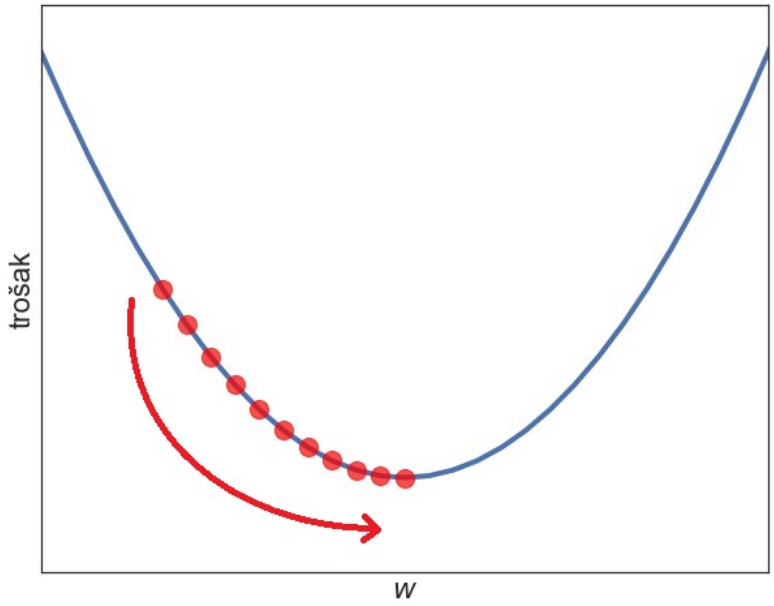
$$\frac{\partial}{\partial w_j}\mathrm{MSE}(\mathbf{w}) = \frac{2}{n}\sum_{i=1}^{n}\left(\mathbf{w}^T\mathbf{x}_i - y_i\right)x_i^{(j)} \tag{3.5}$$

S izračunatim gradijentom je moguće napraviti korak prema minimumu funkcije tako što se gradijent pomnožen s nekom konstantom učenja η oduzme od trenutnog vektora parametara:

$$\mathbf{w}^{(\text{iduci korak})} = \mathbf{w} - \eta\nabla_{\mathbf{w}}\mathrm{MSE}(\mathbf{w}) \tag{3.6}$$

Bitno je pronaći dobru konstantu učenja za pojedinu funkciju. Kod prevelikih iznosa η bit će preveliki koraci i algoritam će preskočiti minimum ili divergirati, dok kod premalih vrijednosti možda neće ni stići do minimuma. Međutim, s idealnom konstantom gradijentni spust izgleda kao što je prikazano na Slici 3.3 gdje se njime minimizira kvadratna funkcija. Ovaj algoritam smanjivanja pogreške može se primijeniti za regresijske i klasifikacijske probleme sve dok je ciljna funkcija konveksna.

Provjera uspješnosti binarnog klasifikatora Kad algoritam prođe kroz proces učenja potrebno je provjeriti njegovu uspješnost. Moramo znati koliko je algoritam koristan u predviđanju s novim podacima. Kako bismo znali da algoritam dobro generalizira, moramo ga provjeriti na podacima na kojima nije učio. Prije svakog projekta strojnog učenja potrebno je razdvojiti podatke na dva dijela: skup podataka za treniranje (engl. *training data*) i skup podataka za testiranje (engl. *test data*) [12]. S obzirom na količinu podataka s kojom raspolažemo, razdvojit ćemo više ili manje podataka, ali najčešće je dovoljno izdvojiti kao skup za testiranje oko 20% podataka. Ako je količina podataka jako velika, može biti dovoljno izdvojiti i 1% podataka [10]. U



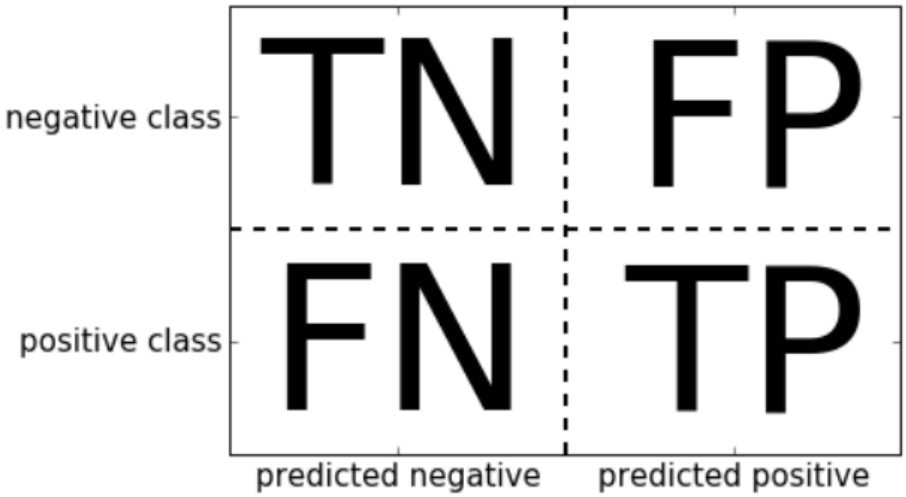
Slika 3.3: Prikaz koraka gradijentnog spusta na kvadratnoj funkciji.

ovom radu radimo s binarnim klasifikatorima koji imaju samo pozitivnu i negativnu klasu u podacima (tj. nešto je ili točno ili nije), te ćemo se fokusirati na njihovu procjenu.

Za procjenu binarnih klasifikatora koristimo nekoliko vrijednosti:

- Istinski pozitiv (engl. *true positive*, TP) – broj predviđenih pozitivnih vrijednosti koje su pozitivne
- Lažni pozitiv (engl. *false positive*, FP) – broj predviđenih pozitivnih vrijednosti koje nisu pozitivne
- Istinski negativ (engl. *true negative*, TN) – broj predviđenih negativnih vrijednosti koje su negativne
- Lažni negativ (engl. *false negative*, FN) – broj predviđenih negativnih vrijednosti koje nisu negativne.

Dobar način za vizualizaciju tih vrijednosti je matrica zbunjenosti (engl. *confusion matrix*) u kojoj svaki redak predstavlja stvarnu klasu, a stupac predviđenu klasu [10]. Matrice zbunjenosti se često koriste za procjenu rezultata klasifikacijskih algoritama [10, 12]. Matrica zbunjenosti prikazana je na Slici 3.4.



Slika 3.4: Matrica zbunjenosti za binarnu klasifikaciju [12].

Pomoću tih vrijednosti moguće je definirati nekoliko mjera kojima se procjenjuje točnost modela [10]:

- Točnost (engl. *accuracy*) – Postotak točno svrstanih podataka:

$$\text{tocnost} = \frac{\text{broj tocno svrstanih predvidjanja}}{\text{ukupan broj predvidjanja}} \tag{3.7}$$

- Preciznost (engl. *precision*) – Točnost pozitivne ili negativne klase. Ako nas zanima koji postotak od svih pozitivno predviđenih podataka je algoritam dobro predvidio računati ćemo sa:

$$\text{preciznost} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.8}$$

  Slično se računa točnost negativno predviđenih podataka.

- Osjetljivost (engl. *recall*) – Predstavlja postotak predviđene klase koji je točno klasificiran. U slučaju predviđanja pozitivne klase on glasi:

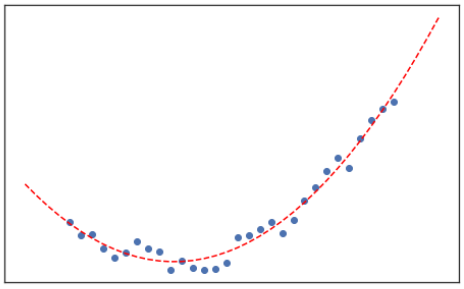
$$\text{osjetljivost} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.9}$$

Očigledno je da želimo imati što veću preciznost i osjetljivost, ali potrebno je odlučiti što nam je najbitnije za određeni problem. U slučaju ovog rada, gdje se nadamo pronaći egzoplanete, bit će nam najbitnija osjetljivost zbog jako malog broja egzoplaneta koje očekujemo. Zbog toga ćemo morati žrtvovati preciznost i dobiti puno više

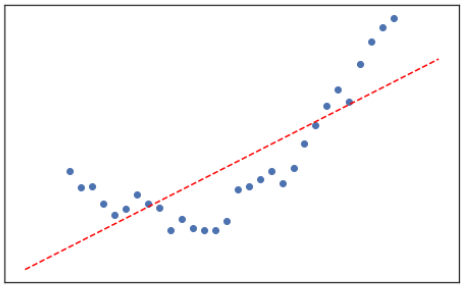
lažnih pozitiva. I u primjeni danas poznatih observacijskih metoda detekcije egzoplaneta teško je biti siguran ima li zvijezda egzoplanet u svojoj okolini. U ovom trenutku (14. 6. 2021.) postoji 4401 potvrđenih egzoplaneta, a čak 6625 kandidata [1]!

Neki problemi u primjeni strojnog učenja Već smo spomenuli kako izborom algoritma biramo i funkciju koja opisuje podatke. Neki algoritmi podržavaju više različitih funkcija. Ako model jako loše predviđa, to može biti znak da je odabrana prejednostavna funkcija za zadani problem. Ako model strojnog učenja nije u stanju opisati fine detalje skupa podataka, onda je došlo do podnaučenosti (engl. *underfitting*). Suprotno, možemo naučiti model da ima odlične rezultate na skupu podataka za testiranje. U tom slučaju može doći do prenaučenosti (engl. *overfitting*) do kojeg dolazi kad je funkcija prekomplicirana za zadani skup podataka [10, 12]. Na Slici 3.5 prikazani su slučajevi podnaučenosti i prenaučenosti za točke koje su generirane kvadratnom funkcijom s nasumičnim šumom. Linearni model je prejednostavan za predviđanje kvadratne funkcije, a polinom 16. stupnja se pak prilagođava podacima predobro. Kod prenaučenosti algoritam će jako dobro predviđati podatke na kojima je učio, ali se ne možemo nadati da će se model dobro generalizirati na skup podataka na kojima nije učio.

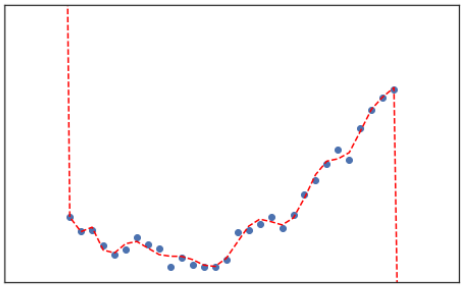
Unakrsna provjera (engl. *cross-validation*) Kako bismo smanjili utjecaj navedenih problema, potrebno je poduzeti specijalne metode provjere. Nije dovoljno testirati model samo na skupu podataka za testiranje, a pogotovo nije dobro vršiti prilagodbu parametara modela kako bi postigao što bolje rezultate na tom skupu. Time smo samo osigurali dobre rezultate na testu, ali ne generalno. Kada bismo tako optimiziran model na podacima za testiranje pustili u produkciju, bilo bi nemoguće predvidjeti hoće li dobro raditi. Zbog toga je korisno izdvojiti skup za provjeru (engl. *validation set*) iz podataka za treniranje. Taj skup se koristi za evaluaciju modela prije nego što se model primijeni na skup za testiranje. Ta metoda zove se metoda izdvajanja (engl. *hold-out method*) [10]. Njen problem je što jako ovisi o podjeli skupa za učenje, što znači da nije poznato hoće li nam rezultati modela biti reprezentabilni u generalnoj situaciji. Kako bismo dobili prosječnu korisnost modela, možemo primijeniti metodu izdvajanja više puta, ali s različitim podskupovima za provjeru. Na ovaj način ne prepuštamo točnost modela slučajnosti odabira podataka za provjeru. Ta



(a)



(b)



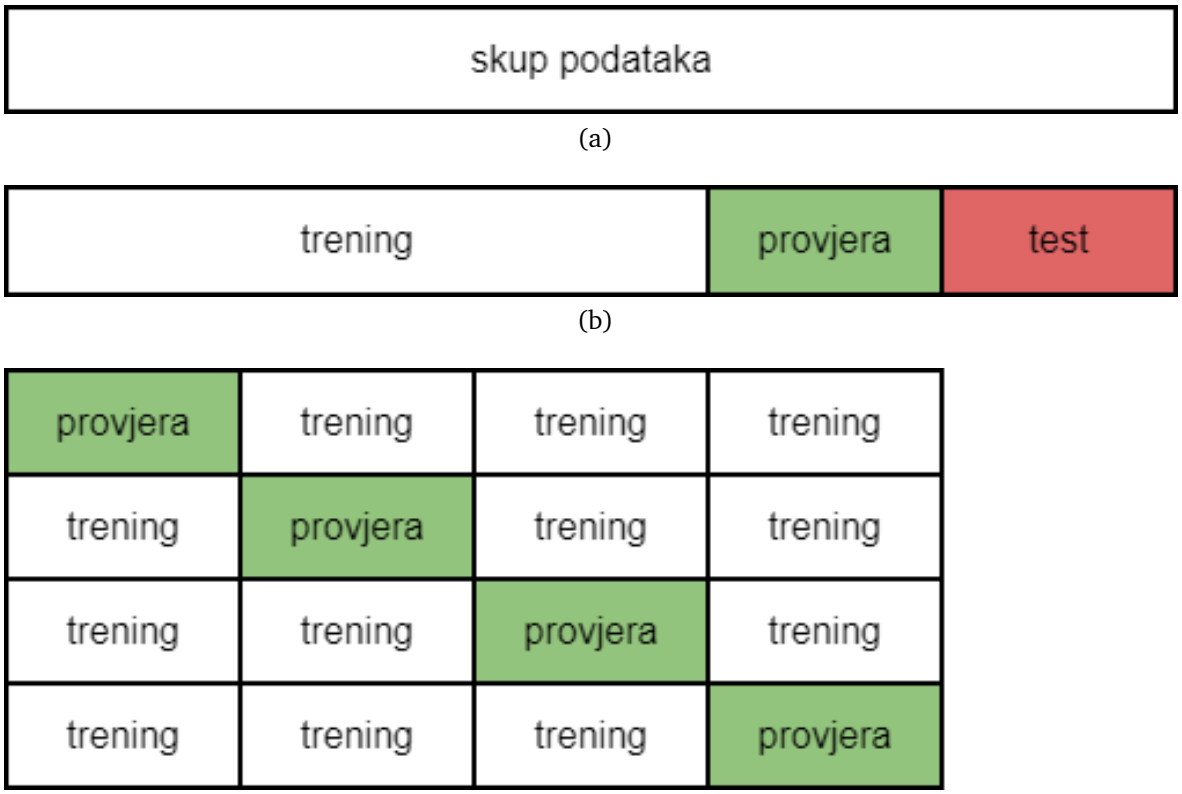
(c)

Slika 3.5: Predviđanje kvadratnom funkcijom (a), linearnom funkcijom (b) i polinomom 16. stupnja (c).

metoda se naziva *k*-struka unakrsna provjera, pri čemu je *k* broj podskupova na koje se dijeli skup podataka za treniranje. U svakoj iteraciji provjere bit će odabran jedan podskup za provjeru, dok će se ostali skupovi koristiti za učenje modela. Obje metode prikazane su za usporedbu na Slici 3.6. Nakon što smo uvjereni da su rezultati dobri, možemo poslati model na pravi i zadnji test na skupu podataka za testiranje koji smo izdvojili na početku. Kada algoritam ima dobre rezultate na skupu podataka za testiranje, a loše na unakrsnoj provjeri, to može biti indikacija da je došlo do prenaučenosti. Ako je model loš i na testnom skupu i na unakrsnoj provjeri, onda je podnaučen [10].

3.3 Logistička regresija

Jedan od modela koji se koriste u ovom radu je logistička regresija. Ona je inačica regresije koja predviđa vjerojatnost nekog događaja na skupu podataka čije su zavisne varijable isključivo binarne, odnosno svaki pojedini podatak pripada jednoj od dvije moguće klase. Klase poprimaju vrijednosti $y = 1$ za pozitivnu i $y = 0$ za negativnu



Slika 3.6: Shematski prikaz skupa podataka (a), metoda izdvajanja (b) i *k*-struka unakrsna provjera (c).

klasu, ili drugim riječima nešto je točno, ili nije točno. Kada je predviđena vjerojatnost za neki podatak veća od 50% logistička regresija klasificirat će je kao pozitivnu klasu, a u suprotom kao negativnu [10]. Za naš skup podataka zvijezde s planetima predstavljaju pozitivnu klasu, a ako nemaju planet u svojoj okolini negativnu klasu.

Funkcija poveznica (engl. *link function*) Logistička regresija jedan je od generaliziranih linearnih modela čiji je zadatak povezati značajke $(x_1, x_2, \ldots, x_m)$ s vjerojatnosti p kojom se predviđa pripadnost podatka pojedinoj klasi. Modeli koji se koriste za klasificiranje binarnih podataka, kao što to radi logistička regresija, preslikavaju linearnu kombinaciju značajki na interval $[0, 1]$. To se ostvaruje transformacijom vjerojatnosti pomoću funkcije poveznice $g(p)$ koja preslikava interval $[0, 1]$ na skup realnih brojeva [13] kako bi se dobila linearna veza:

$$g(p_i) = \sum_{j=1}^{m} x_{ij} \mathrm{w}_j;\ i = 1, \ldots, n. \tag{3.10}$$

Odnosno u vektorskom obliku moguće je obuhvatiti sva predviđanja:

$$g(\mathbf{p}) = \mathbf{X}^T\mathbf{w}. \tag{3.11}$$

Logistička funkcija koristi *logit* funkciju kao funkciju poveznicu [13]:

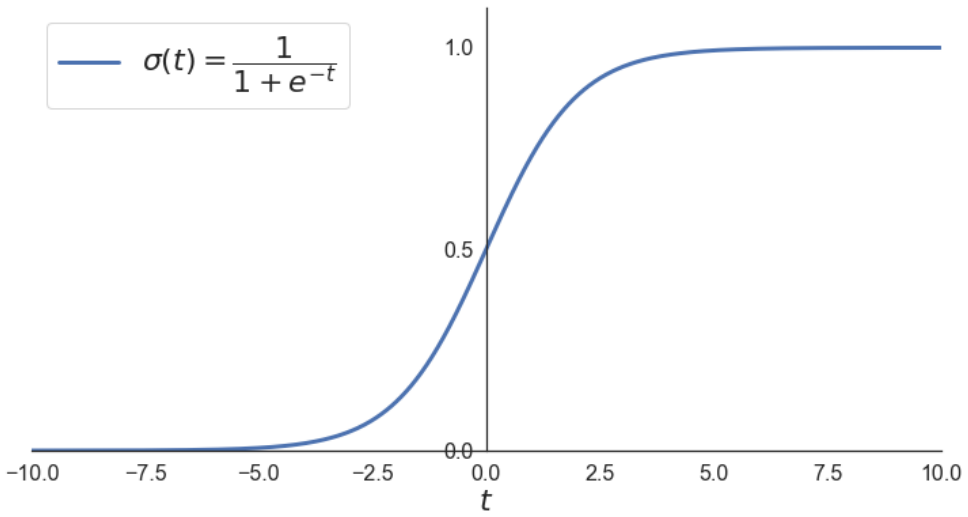
$$g(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{3.12}$$

Uvrštavanjem funkcije 3.12 u jednadžbu 3.11 i rješavanjem dobije se funkcija vjerojatnosti koja ovisi o ulaznim značajkama $\mathbf{X}$, a parametrizirana je težinskim koeficijentima $\mathbf{w}$:

$$p(\mathbf{X};\mathbf{w}) = \frac{1}{1+e^{-(\mathbf{X^T w})}}. \tag{3.13}$$

Jednadžba 3.13 zove se logistička funkcija. Ta funkcija, za razliku od 3.12 koja preslikava vjerojatnost na skup realnih brojeva, preslikava linearnu funkciju sa skupa realnih brojeva na interval $[0, 1]$. Prepoznatljiva je po specifičnom S-obliku te pripada skupu krivulja koje se nazivaju sigmoide (vidi Sliku 3.7). Njen pojednostavljen izraz je [10]:

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{3.14}$$



Slika 3.7: Logistička funkcija.

Predviđanje logističke regresije Logistička regresija računa otežanu sumu ulaznih značajki, kao što to radi linearna regresija, ali je koristi kao argument logističke

funkcije koja daje vrijednosti između 0 i 1. Hipoteza $h(\mathbf{x})$ logističke regresije je upravo logistička funkcija koja se prilagođava podacima i njome se nakon učenja predviđa vjerojatnost $\hat{p} = h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{x}^T\mathbf{w})$ da podatak $\mathbf{x}$ pripada pozitivnoj klasi [10]. Predviđene vrijednosti $\hat{y}$ lako se dobivaju odabirom granice odlučivanja $\hat{p} = 0.5$:

$$\hat{y} = \begin{cases} 0, & \text{ako } \hat{p} \leq 0.5 \\ 1, & \text{ako } \hat{p} > 0.5 \end{cases} \tag{3.15}$$

Granicu je također moguće proizvoljno odabrati ako se time može povećati točnost modela. Nije zagarantirano da se krivulja najbolje prilagodila podacima nakon učenja pa se uvijek može provjeriti donosi li neka druga granica bolje rezultate. U generalnom slučaju veća granica za odluku znači da zahtjevamo veću sigurnost modela za pozitivnu klasu.

Funkcija troška i učenje Logistička regresija uči pomoću gradijentnog spusta na sličan način kao što je opisano u poglavlju 3.2 na primjeru kvadratne funkcije. Funkcija troška za logističku funkciju ovisi o klasi predviđenog podatka, definirana je jednadžbom 3.16 i prikazana je na Slici 3.8 [10]. Na istoj slici vidljivo je na plavoj krivulji da funkcija troška ima velike vrijednosti kad logistička regresija predviđa male vjerojatnosti $\hat{p}$ za pozitivnu klasu. Što je predviđena vjerojatnost veća, funkcija troška je manja, a predviđanje sve točnije. U suprotnom slučaju narančasta krivulja, koja opisuje predviđanje za negativnu klasu, pokazuje da funkcija troška poprima velike vrijednosti kad vjerojatnost $\hat{p}$ raste.

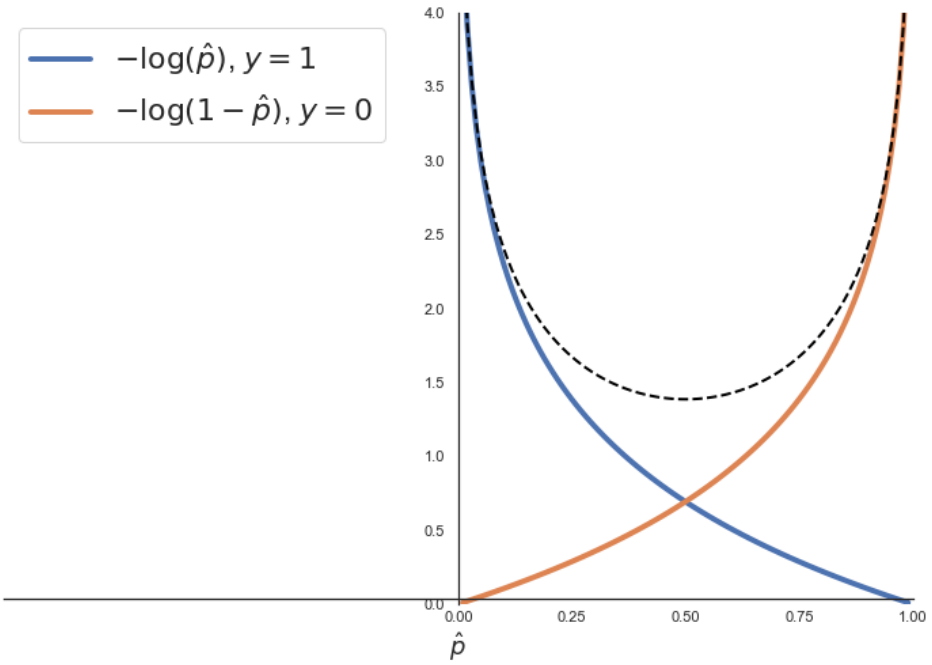
$$\text{trošak}(\mathbf{w}) = \begin{cases} -\log(\hat{p}), & \text{ako } y = 1 \\ -\log(1-\hat{p}), & \text{ako } y = 0 \end{cases} \tag{3.16}$$

Funkciju troška moguće je objediniti u jedan izraz prikazan u jednadžbi 3.17. Nakon toga je potrebno izračunati parcijalnu derivaciju funkcije troška (3.18) kako bi se mogao primijeniti gradijentni spust za njenu minimizaciju [10]. Za $y_i = 1$ u 3.17 u zagradi ostaje samo član $\log(\hat{p}_i)$, dok će za $y_i = 0$ ostati samo $\log(1 - \hat{p}_i)$ kao što je zadano u 3.16. S obzirom da funkcija troška uzima u obzir cijeli skup podataka, podijeljena je s brojem podataka n, te predstavlja srednju vrijednost troška. Pojedina parcijalna derivacija iz 3.18 predstavlja usrednjenu pogrešku predviđanja pomnoženu

sa j-tom značajkom podatka. Računom parcijalnih derivacija svih značajki ispuni se vektor gradijenta nakon čega je moguće izračunati nove vrijednosti težinskih koeficijenata prema izrazu 3.6 sve dok gradijentni spust ne stigne do minimuma funkcije troška 3.17.

$$J(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right] \tag{3.17}$$

$$\frac{\partial}{\partial \mathrm{w}_j} J(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\sigma(\mathbf{w}^T\mathbf{x}_i) - y_i\right) x_i^{(j)} \tag{3.18}$$



Slika 3.8: Funkcija troška logističke funkcije.

3.4 Stabla odlučivanja

Stabla odlučivanja (engl. *decision trees*) su široko primjenjiv algoritam strojnog učenja. Mogu se koristiti za regresiju i klasifikaciju podataka. Stabla uče postavljanjem niza pitanja o značajkama kako bi došlo do odluke [12]. Za binarnu klasifikaciju pitanja imaju samo dva moguća odgovora; neki uvjet vrijedi ili ne vrijedi te je i stablo koje će se izgraditi binarno. To znači da se prilikom svakog pitanja podaci razgranaju na dva dijela te se na taj način gradi stablo. Osim što su stabla vrlo brz algoritam, velika

prednost im je što su rezultati jednostavni za interpretirati i lako se mogu usporediti sa svakidašnjim odlukama. Kada bismo odlučivali je li potrebno ponijeti jaknu kada izlazimo vjerojatno bismo se prvo zapitali je li vani sunčano, ili kišno vrijeme. Jakna će nam također biti potrebna ako je vani niska temperatura. Jedno moguće stablo za navedenu odluku prikazano je na Slici 3.9. Prvi čvor stabla koji se nalazi na vrhu naziva se korijen i u njemu su sadržane sve ulazne značajke. Korijen se grana na čvorove odluka koji klasificiraju vrijednosti ulaznih podataka sve dok ne stignu do kranjih čvorova tj. listova koji predstavljaju izlazne podatke stabla, odnosno konačnu odluku. Svakom stablu moguće je odrediti dubinu stabla brojeći najveći broj grana od korijena do lista. Dubina stabla u primjeru na Slici 3.9 je dva. Na istom primjeru vidljiv je jedan problem stabala odlučivanja; na desnoj strani stabla ono odlučuje da jakna nije potrebna ako ne puše jak vjetar za vrijeme kiše. Sasvim je moguće da je temperatura niska, a stablo tu mogućnost nije obuhvatilo. To se dogodilo zato što stablo na Slici 3.9 odlučuje samo na temelju jedne značajke u svakoj iteraciji, odnosno postavlja samo jedno pitanje za svaku odluku. Kod kompliciranijih primjera stabla mogu odlučivati na temelju više značajki istovremeno prilikom svake odluke, a moguće je i kontrolirati broj značajki koji se uzima u obzir prilikom čega se odabir značajki vrši nasumično, ili prema nekoj mjeri troška. Osim broja značajki bitno je posvetiti pažnju i dubini stabla. Preduboka stabla neizbježno su prenaučena jer postavljaju pitanja i granaju se sve dok savršeno ne opisuju podatke za treniranje. U ovom radu stabla odlučivanja koriste se za klasifikaciju zvijezda s planetom i bez planeta, a značajke na temelju kojih stablo odlučuje su intenziteti svjetlosti zvijezda u pojedinom trenutku.

Funkcija troška i učenje Jedan od načina kojim stablo odlučuje kako razgranati odluku je mjera čistoće koja se zove gini nečistoća (engl. *gini impurity*) [10]:

$$G^{(i)} = 1 - \sum_{k=1}^{n} \left(p_k^{(i)} \right)^2 \tag{3.19}$$

pri čemu je $p_k^{(i)}$ omjer broja podataka koji pripadaju klasi k i ukupnog broja podataka u i-tom čvoru stabla. Prema gini nečistoći čvor je čišći što je njena vrijednost manja. Potpuno čisti čvor imati će vrijednost $G^{(i)} = 0$, a to znači da sadrži samo podatke koji pripadaju istoj klasi. Suprotno, kada čvor sadrži podatke koji pripadaju različitim

Slika 3.9: Stablo odlučivanja koje odlučuje je li potrebno ponijeti jaknu.

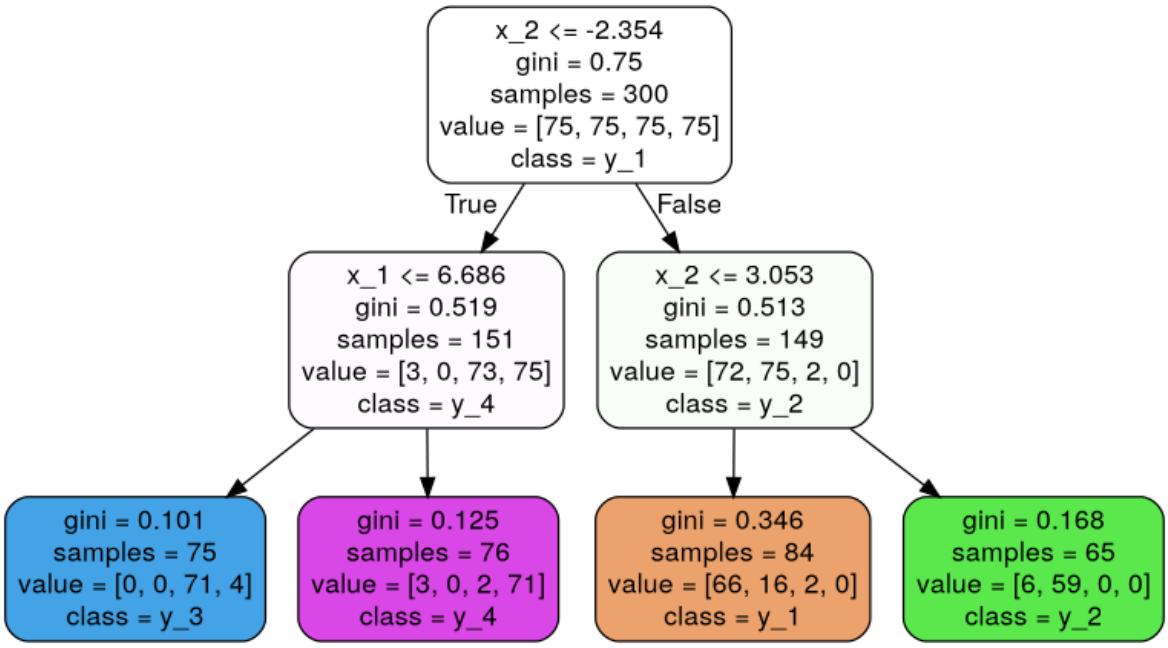
klasama, on će imati nečistoću različitu od 0. Na primjer, čvor na dubini 1 lijevo na Slici 3.10a ima gini nečistoću $G = 1 - \frac{3}{151} - \frac{0}{151} - \frac{73}{151} - \frac{75}{151} = 0.519$. Funkcija troška koja se koristi u Scikit-Learn paketu za tu svrhu zadana je jednadžbom s pomoću koje se računa trošak značajki k s graničnom vrijednosti t_k po kojoj će se podijeliti podaci [10]:

$$J(k, t_{\text{k}}) = p_{\text{lijevo}} G_{\text{lijevo}} + p_{\text{desno}} G_{\text{desno}} \tag{3.20}$$

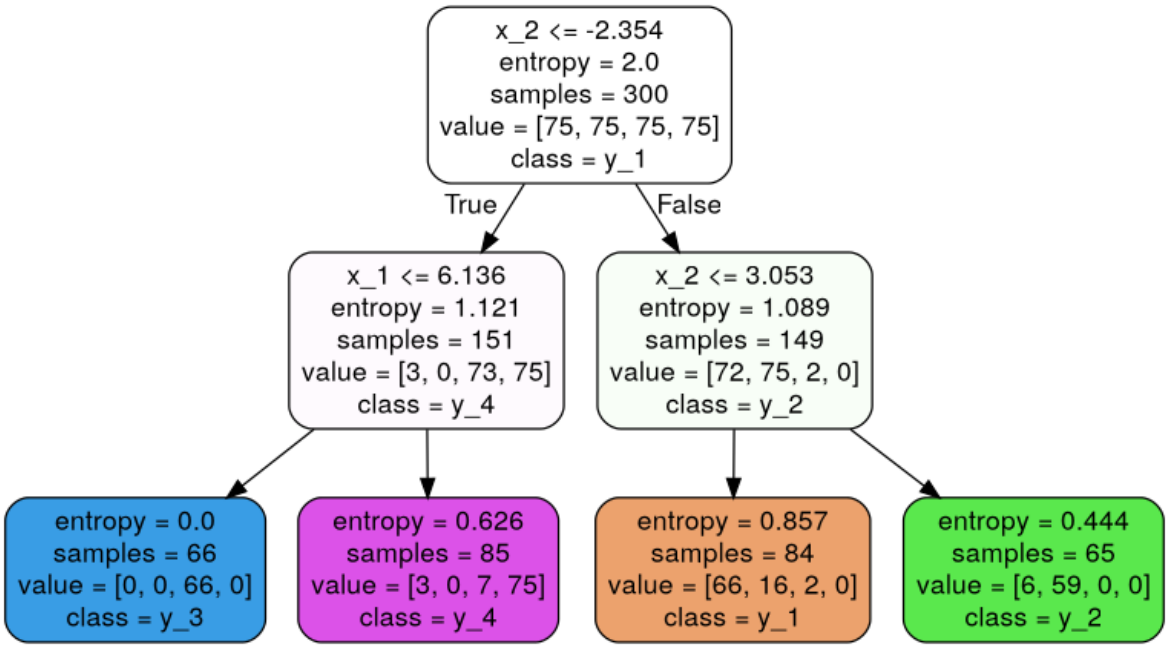
Minimizacijom funkcije 3.20 pronalazi se najbolja značajka i njena vrijednost za iduće grananje. Cilj je postići najmanju vrijednost nečistoće u idućem grananju računajući sve moguće vrijednosti nečistoće za lijevu i desnu granu. Na sličan način moguće je učiti stabla odlučivanja primjenom entropije kao mjere nečistoće [10]:

$$H^{(i)} = -\sum_{k=1}^{n} p_k^{(i)} \log_2 \left(p_k^{(i)} \right). \tag{3.21}$$

Entropija je koncept iz termodinamike koji predstavlja mjeru neuređenosti sustava. U korijenu stabla, kad još podaci nisu razvrstani i stablo nije naučeno, entropija svih podataka je maksimalna. Idealan čvor bi imao entropiju jednaku 0 što bi značilo da je potpuno uređen, odnosno da sadrži samo podatke koji pripadaju istoj klasi. Na primjer, čvor na dubini 2 desno na Slici 3.10b ima entropiju $H = -\frac{6}{65} \log_2 \left(\frac{6}{65} \right) - \frac{59}{65} \log_2 \left(\frac{59}{65} \right) - \frac{0}{65} \log_2 \left(\frac{0}{65} \right) - \frac{0}{65} \log_2 \left(\frac{0}{65} \right) = 0.444$.



(a)

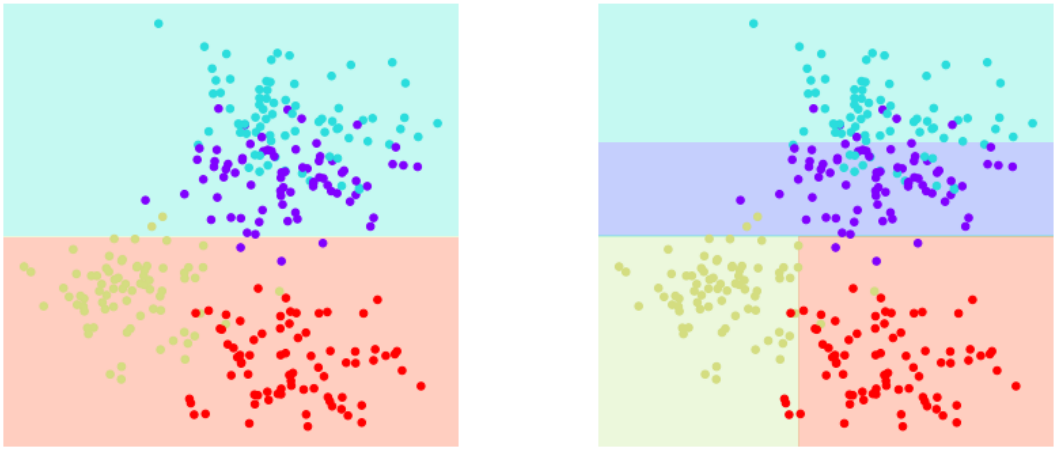


(b)

Slika 3.10: Primjer stabla odlučivanja s gini nečistoćom (a) i entropijom (b) kao mjerom čistoće čvora.

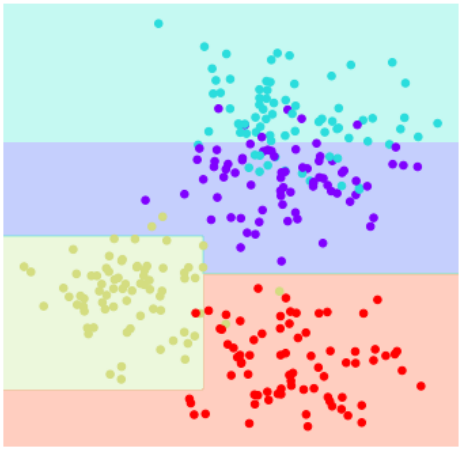
Problemi stabla odlučivanja Stabla odlučivanja su izrazito dobra u prilagođavanju na zadani skup podataka ako im se ne ograniči maksimalna dubina stabla. U tom slučaju čvorovi će se granati sve dok svaki list ne bude potpuno čist po mjeri nečistoće i zbog toga su stabla odlučivanja sklona prenaučenosti [12]. Prilikom učenja stabla važno je obratiti pozornost na maksimalnu dubinu i ograničiti ju u trenutku kada dubina stabla više ne povećava značajno njegovu preciznost. Na Slici 3.11 prikazano

je stablo koje kategorizira četiri skupine podataka sa različitim maksimalnim dubinama. Za dubinu stabla jednaku jedan (kao na Slici 3.11a) stablo je uspjelo razdvojiti podatke samo na plave i crvene točke. Klasifikacija će biti sasvim točna za te podatke, ali ne možemo dobiti nikakvu informaciju o žutim i ljubičastim. U ovom slučaju stablo je prejednostavno za zadani skup podataka pa je podnaučeno. Povećavanjem dubine (kao što je prikazano na Slikama 3.11b i 3.11c) stablo puno bolje opisuje zadane podatke. Na Slici 3.11d prikazana je klasifikacija stabla koje je prenaučeno zbog prevelike maksimalne dubine. Žuto područje zadire u crveno zbog jedne žute točke, koja u nekoj realnoj situaciji može biti rezultat šuma u nekom mjerenju, te se pojavljuju plave linije u ljubičastom području. Razlog tome je što se stablo potpuno prilagodilo podacima za učenje i ono neće dobro generalizirati nove podatke. Za prikazane podatke bilo bi dovoljno odabrati dubinu stabla $n = 2$ ili $n = 3$.

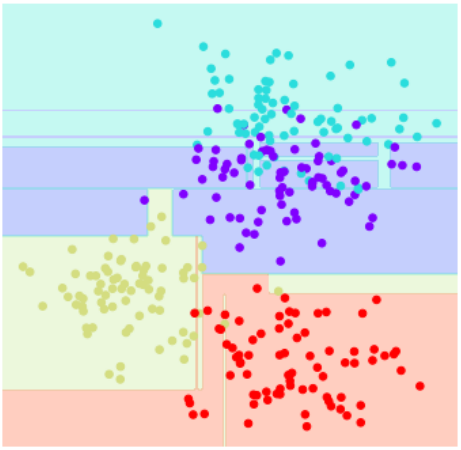


(a)

(b)



(c)



(d)

Slika 3.11: Rezultat klasifikacije točaka pomoću stabla odlučivanja maksimalne dubine $n = 1$ (a), $n = 2$ (b), $n = 3$ (c) i $n = 10$ (d). Točke su nasumično generirane oko četiri nasumično odabrana središta gausijanskom raspodjelom s obzirom na udaljenost od pojedinog središta. Svaka boja pripada jednom od četiri središta, a regije pripadnih boja predstavljaju predviđanje stabla odlučivanja.

4 Potraga za egzoplanetima primjenom strojnog učenja

4.1 Korišteni alati i metode

Python Python je objektno orijentirani programski jezik visoke razine koji svoju popularnost može zahvaliti jednostavnoj sintaksi, te licenci otvorenog koda (engl. open source) zbog čega postoji mnoštvo besplatnih modula [14]. Zbog jednostavnosti i pristupačnosti Python je zanimljiv, ne samo osobama koje se bave informatikom, već i znanstvenicima koji se primarno ne bave programiranjem. Glavna prednost Pythona je brzina izrade koda, ali ne i brzina izvršavanja programa s obzirom da je interpretiran jezik. Taj problem je djelomično zaobiđen implementiranjem i kompajliranjem modula u programskim jezicima C i C++. Zbog toga je Python posebno koristan za usko specijalizirane zadatke koji koriste takve module jer neće biti značajne razlike u vremenu izvršavanja u usporedbi s jezicima koji se kompajliraju u strojni kod. Python je posebno značajan u području strojnog učenja i analize podataka. Svoju snagu također pokazuje kada je potrebno na brzinu napisati probni ili pokazni program te je Python idealan i za iskusne programere i za početnike. Također je sve češći odabir za uvod u programiranje u hrvatskim školama.

Jupyter bilježnica Jupyter bilježnica je web aplikacija otvorenog koda koja je dio projekta Jupyter [15]. Projekt podržava nekoliko programskih jezika među kojima je i Python. Bilježnice se koriste za sekvencijalno pisanje koda, izvršavanje te vizualizaciju podataka u istom programskom okruženju. U Jupyterovim bilježnicama je jednostavno dijeliti kod, dokumente i rezultate u mnoštvu formata. To ih čini izvrsnim alatom za podatkovne znanosti (engl. data science) zbog toga što su vrlo jednostavne za analizu podataka i strojno učenje.

NumPy NumPy je jedan od Pythonovih modula koji se koristi za računanje u znanosti [16]. Veći dio koda u NumPy modulu je pisan u programskom jeziku C što njegove funkcije čini vrlo brzim i efikasnim. Implementirano je mnoštvo matematičkih objekata kao što su vektori i matrice te rutina kao što su Fourierove transformacije, generiranje nasumičnih brojeva itd. Rad s vektrima i matricama je puno brži od standardnih Pythonovih lista.

Scikit-learn Scikit-learn jedan je od popularnih modula u svijetu podatkovne znanosti [17]. U modulu su implementirani algoritmi za strojno učenje koji se jednostavno mogu primijeniti u sklopu istog sučelja. Scikit-learn je pisan na temelju NumPy i SciPy modula u kojima je većina funkcija pisana u programskim jezicima C i Fortran.

Pandas Najpoznatiji modul za rad u podatkovnoj znanosti je Pandas [18]. Brz je i efikasan u radu s podacima koji su definirani kao DataFrame objekti. Podaci su spremljeni kao tablice. Implementirano je puno metoda za rad s podacima. Te metode su optimizirane i pisane su u programskim jezicima C i Cython. Zbog toga je rad s velikom količinom podataka jednostavan i intuitivan. Podatci se mogu čitati i spremiti u različitim formatima kao što su CSV, Microsoft Excel, ili SQL baza podataka.

Matplotlib i seaborn Matplotlib je sveobuhvatan modul za vizualizaciju podataka i izradu animacija [19]. Sve karakteristike izrađenih dijagrama moguće je mijenjati i spremiti kao datoteke u nekoliko formata. Seaborn je razvijen na bazi matplotlib modula i specijaliziran je za vizualizaciju statističkih podataka [20]. Njegova prednost nad matplotlib modulom je direktna podrška za rad s pandas modulom što značajno olakšava vizualizaciju podataka iz pandas DataFrame objekta.

4.2 Skup podataka

Podaci korišteni u ovom radu dio su treće kampanje druge Kepler misije nazvane po svemirskom teleskopu Kepler koji je NASA razvila u svrhu potrage za egzoplanetima. Misija se fokusirala na snimanje zvijezda u regiji Mliječne staze u kojoj se nalazi Zemlja, s ciljem pronalaska sličnih planeta koji bi potencijalno podržavali život. Svaka kampanja je zbog putanje teleskopa i kuta upada Sunca bila ograničena na snimanje zvijezda oko 80 dana [2]. Nakon micanja šuma u signalu, prikupljene podatke NASA objavljuje u arhivima koji su besplatno dostupni svima [21, 22]. Skup podataka koji se koristi u radu preuzet je sa Kaggle web stranice na kojoj je korisnik izvdojio intenzitete svjetlosti zvijezda treće kampanje zajedno sa svim potvrđenim planetima ostalih kampanja iz Mikulski arhive [23]. U tablici se nalazi 5657 redaka koji predstavljaju snimanje intenziteta svjetlosti pojedine zvijezde u periodu od 80 dana. U stupcima koji predstavljaju značajke zapisano je 3197 intenziteta svjetlosti svake zvijezde koji su snimljeni u jednakim vremenskim razmacima. Dio podataka prikazan

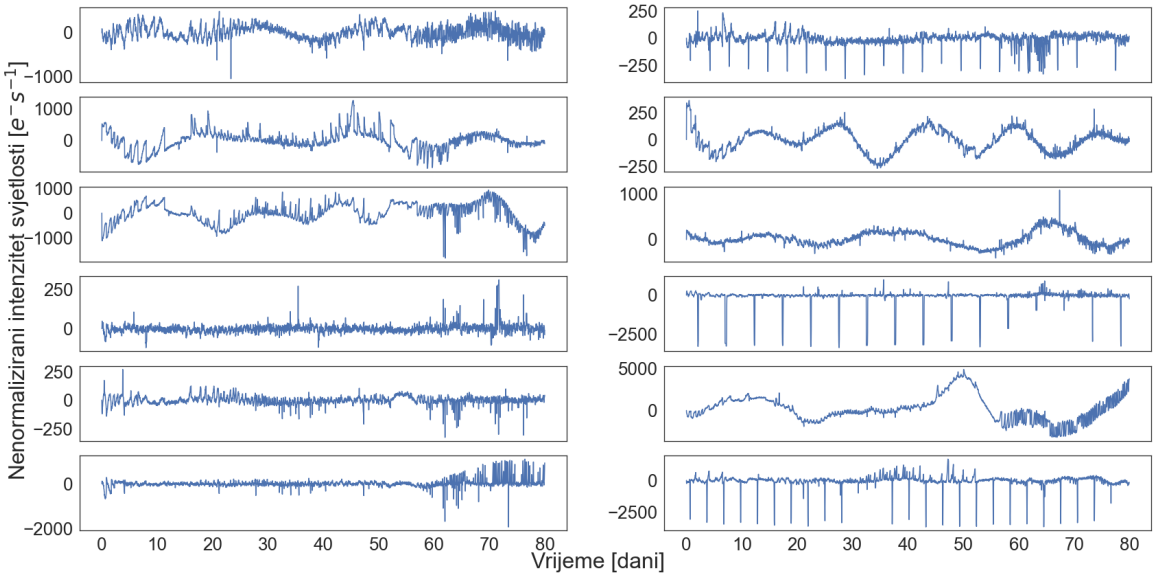
	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	...
0	2	93.85	83.81	20.10	-26.98	...
1	2	-38.88	-33.83	-58.54	-40.09	...
2	2	532.64	535.92	513.73	496.92	...
...	...	...	...	...	...	...
5082	1	-91.91	-92.97	-78.76	-97.33	...
5083	1	989.75	891.01	908.53	851.83	...
5084	1	273.39	278.00	261.73	236.99	...
...	...	...	...	...	...	...

Tablica 4.1: Isječak tablice intenziteta svjetlosti zvijezda.

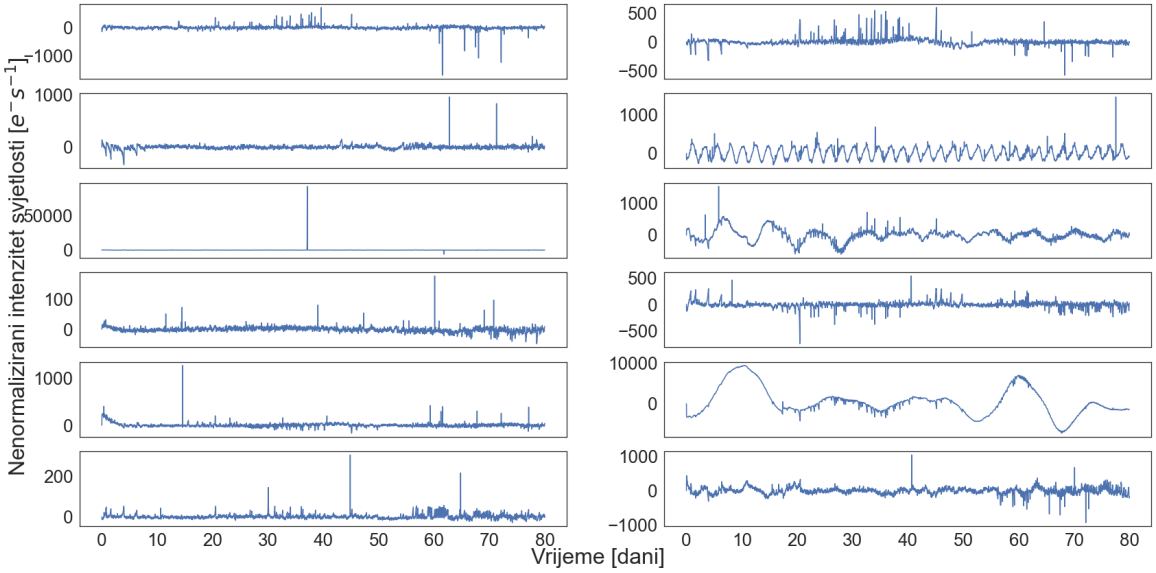
je u Tablici 4.1. Jedan stupac (LABEL u Tablici 4.1) sadrži oznake koje pokazuju ima li zvijezda planet u svojoj okolini (LABEL=2), ili ga nema (LABEL=1). U tablici se nalazi ukupno 42 potvrđena egzoplaneta, a podijeljena je na skup podataka za treniranje koji sadrži 37 potvrđenih planeta i 5087 zvijezda bez planeta, dok je u skupu za testiranje 5 zvijezda s planetom i 565 bez.

Proučavanje podataka Nekoliko algoritama strojnog učenja testirano je na originalnom skupu podataka. Pokazalo se da ti algoritmi nisu točno predvidjeli niti jedan egzoplanet. Intenziteti svjetlosti nekih zvijezda s planetima i bez njih prikazani su na Slici 4.1. Na slici se vidi da su zvijezde s planetima karakterizirane udubinama u grafu, odnosno padom intenziteta zbog prolaska planeta ispred zvijezde. Šum izgleda periodično i često je sinusoidalan kao što se vidi na zumiranoj Slici 4.2 jednog od signala. Bez daljnje obrade podataka algoritmi strojnog učenja koje koristimo nisu u stanju prepoznati strukturu podataka i udubine u grafu. Gledajući vrijednosti na grafovima očito je da podaci nisu normalizirani, a neki od njih sadrže vrijednosti koje značajno odstupaju od srednje vrijednosti što se najbolje vidi na trećem, petom i još nekoliko grafova sa Slike 4.1b. Uzrok takvim odstupanjima mogu biti nepravilnosti instrumenta i razni izvori šuma, a potrebno ih se riješiti jer će uzrokovati probleme za algoritme strojnog učenja. Na Slici 4.3 prikazana su neka osnovna obilježja skupa podataka. Srednje vrijednosti i medijani se nalaze oko nule, vjerojatno kao rezultat obrade podataka prije njihove objave u arhivu. Gledajući najveće i najmanje intenzitete svjetlosti, te standardne devijacije pojedinih redaka (zvijezda) jasno se vidi da postoje velika odstupanja. Prije obrade signala izbačeni su svi retci kojima su vrijednosti veće od tri standardne devijacije te je skup podataka za treniranje reduciran na 5010 zvijezda. Detaljnije ćemo pratiti transformacije na signalu jedne zvijezde koja

sadrži planet u svojoj okolini. Signal je prikazan na Slici 4.4a. Uske udubine u grafu uzrokovane su prolaskom planeta ispred zvijezde, a između udubina vidljiv je šum koji prati zajednički obrazac.



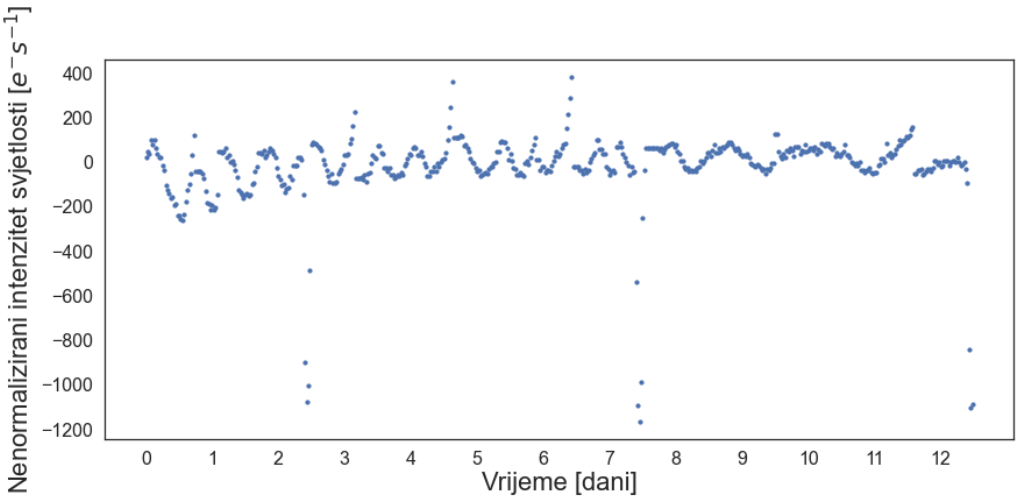
(a)



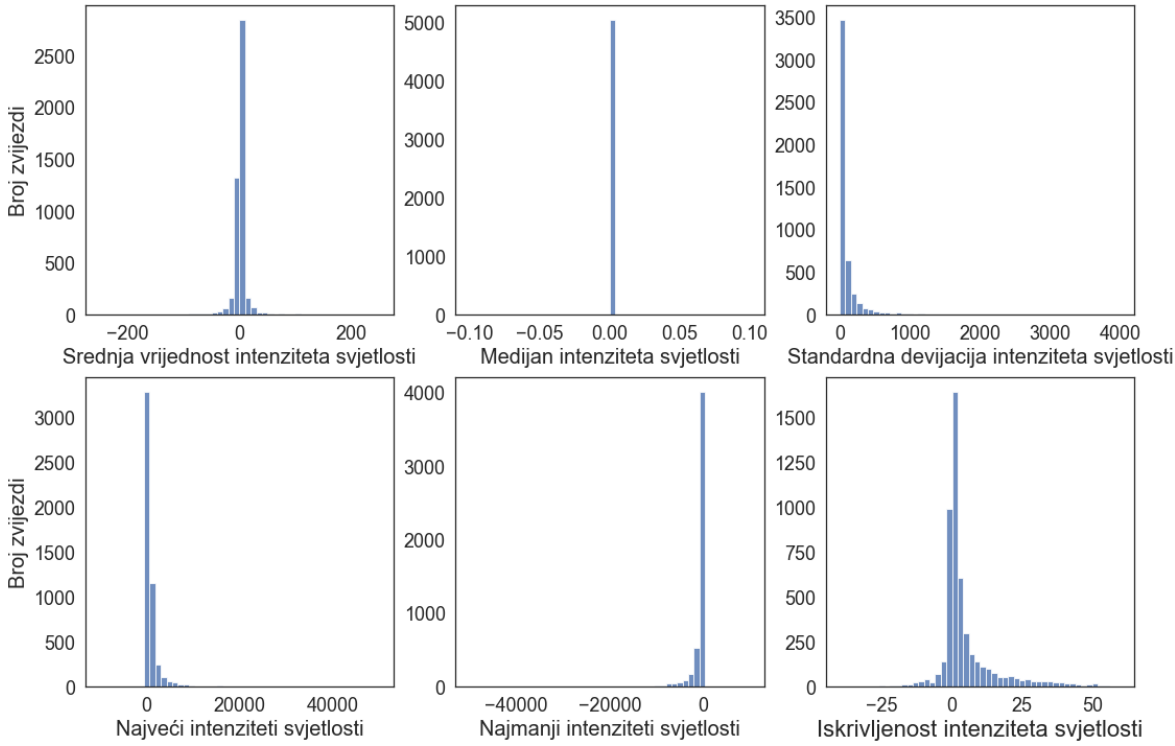
(b)

Slika 4.1: Intenziteti svjetlosti nekih zvijezda s egzoplanetima (a) i bez egzoplaneta (b).

Skaliranje podataka Skaliranje podataka je jako bitno u strojnom učenju jer većina algoritama ne daje dobre rezultate kada se vrijednosti značajki jako razlikuju [10]. Zbog načina na koji radi minimizacija funkcije troška logističke regresije prevelike vrijednosti mogle bi uzrokovati velike promjene koeficijenata prilikom gradijentnog



Slika 4.2: Zumirana slika jedne zvijezde s egzoplanetom.

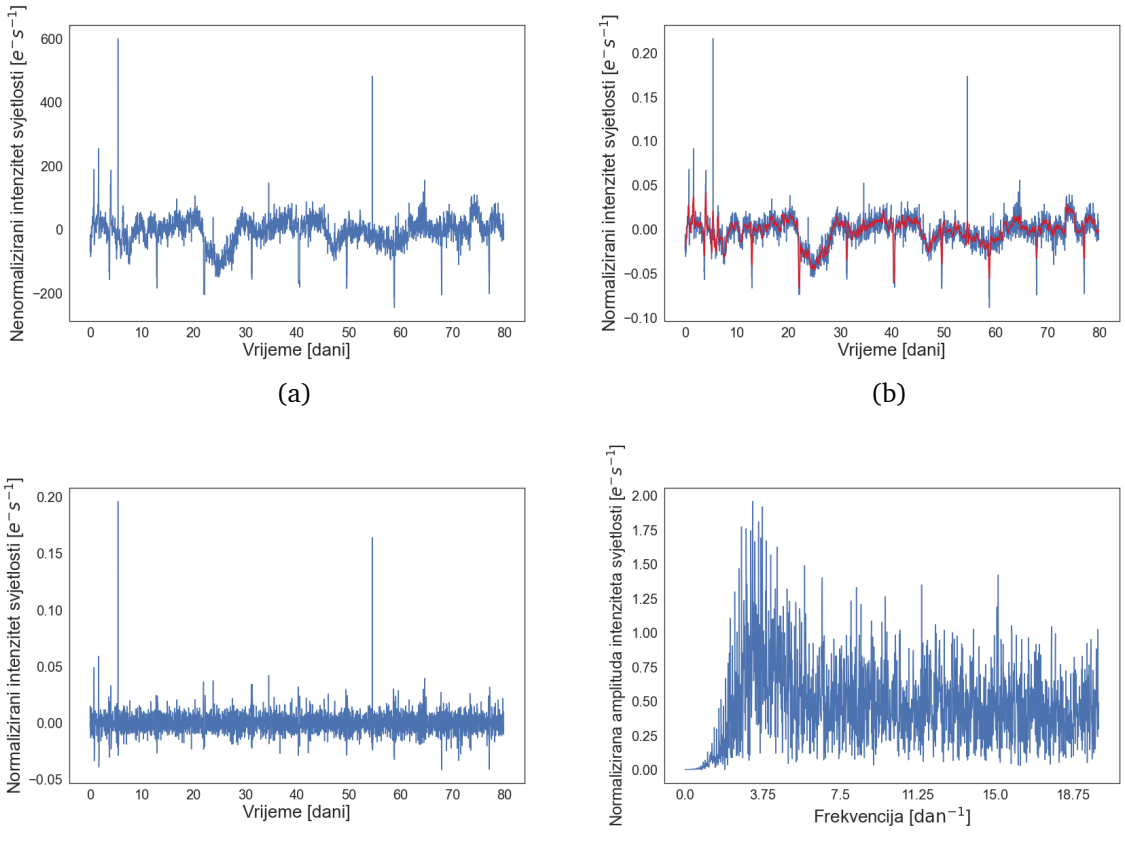


Slika 4.3: Neke osnovne karakteristike skupa podataka.

spusta koji je opisan u odjeljku 3.3. Nije poželjno uzrokovati velike promjene koeficijenata, posebno ako značajka u pitanju ne bi trebala imati velik utjecaj na rezultat, a zbog svoje prevelike vrijednosti ima. Podaci su nakon primjene Savitzky-Golayevog filtra [24] normalizirani, t.j. skalirani tako da se sve vrijednosti nalaze u intervalu $[0, 1]$ koristeći izraz $x_{skalirani} = \frac{x - \min(x)}{\max(x) - \min(x)}$. Na kraju svih transformacija podaci su standardizirani, to jest skalirani tako da im je srednja vrijednost jednaka nuli i standardna devijacija $\sigma = 1$ računajući z-vrijednost $z = \frac{x - \overline{x}}{\sigma}$. Stabla odlučivanja za razliku od logističke regresije nisu osjetljiva na skalu podataka te za njih ovaj korak

nije važan [10].

Savitzky–Golayev filtar S obzirom da će podaci biti prebačeni u frekventnu domenu poželjno je izgladiti šum kako bi se mogli fokusirati na udubine u signalu. Kako bi to postigli primijenili smo Savitzky–Golayev filtar koji se koristi za izglađivanje podataka lokalnom uporabom metode najmanjih kvadrata [24]. Moguće je zadati širinu intervala na kojem se računa i red polinoma koji se prilagođava na podatke u zadanom intervalu. Rezultat filtra prikazan je na Slici 4.4b crvenom bojom. Izglađeni signal oduzimamo od originalnog signala kako bi ostale istaknute samo udubine u grafu koje nas zanimaju, a rezultat toga vidi se na Slici 4.4c. Amplitude udubina se smanjuju zbog interpolacije s okolnim šumom, ali i dalje ostaju prepoznatljive te će Fourierova transformacija moći prepoznati njihovu periodičnost.

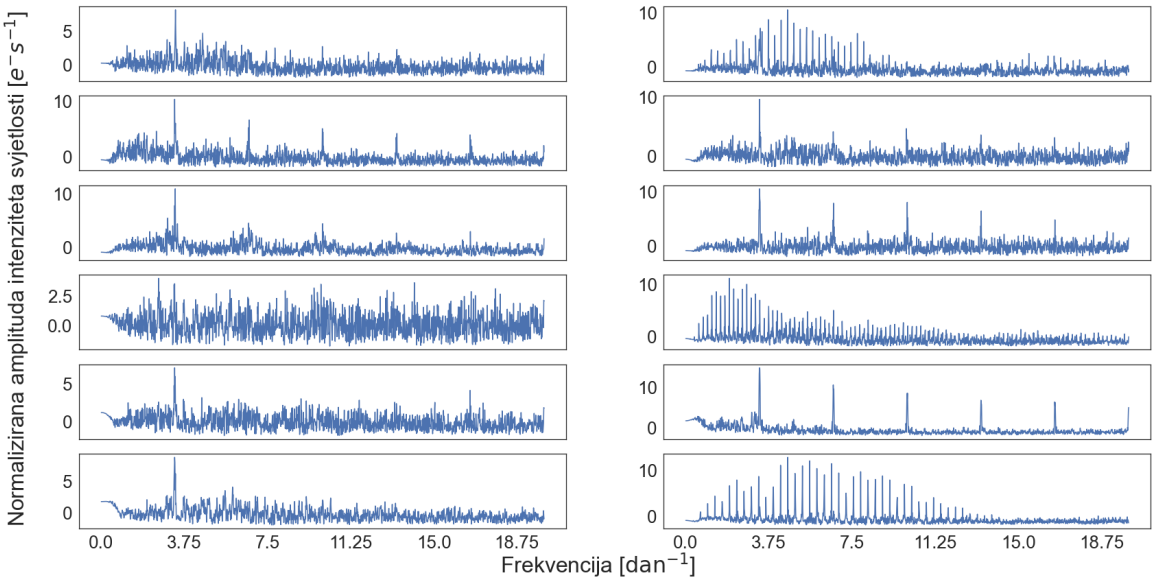


Slika 4.4: Originalni signal (a), Savitzky–Golayev filtar (crveno) vršen nad normaliziranim signalom (plavo) (b), signal nakon oduzimanja filtra od normaliziranog signala (c) i spektar dobiven Fourierovom transformacijom (d).

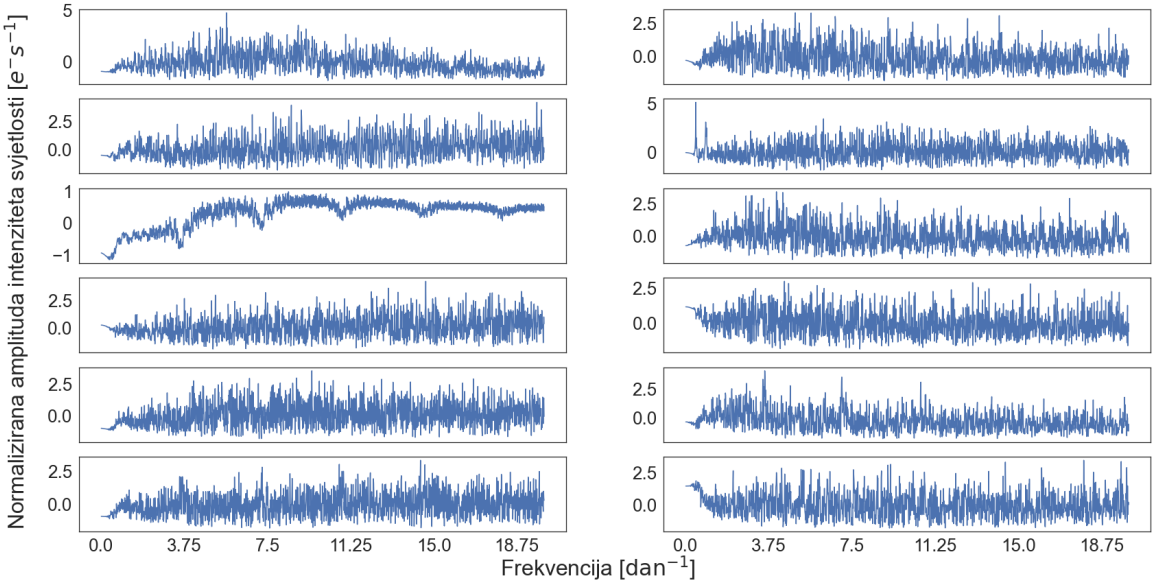
Fourierova transformacija podataka Fourierova transformacija razlaže podatke iz originalne domene u frekventnu, odnosno originalnu funkciju transformira u sumu komponenata koje su u njoj periodične. Transformacija iz funkcije prepoznaje frekvencije svih periodičnih elemenata, a točnost je ograničena principom neodređenosti. S povećanjem mjerenog signala frekventna rezolucija je oštrija, ali na račun rezolucije u vremenskoj domeni jer u dužem uzorku signala postoji više interferencija među periodičnim komponentama. S obzirom da su zvijezde snimane 80 dana i u većini se prolazak planeta događa više puta, pretpostavljamo da je signal reprezentativan i frekventna rezolucija dovoljno oštra za potrebe strojnog učenja. Razlaganjem signala na frekvencije očekujemo kod zvijezda s egzoplanetom vidjeti vrhove među niskim frekvencijama zato što bi prijelazi planeta trebali biti najrjeđi događaji u signalu pored visokofrekventnog šuma. S prethodnim micanjem trenda signala pomoću Savitzky-Golayevog filtra očekujemo da je vjerojatnije da niske frekvencije predstavljaju upravo prolaske planeta. Signal nakon primjene Fourierove transformacije i skaliranja prikazan je na Slici 4.4d na kojoj se jasno vidi da su vrhovi veći na niskim frekvencijama. Kod nekih zvijezda postoje i dodatni vrhovi koji bi mogli biti znak da ima više planeta. Nakon kompletne obrade podataka na Slici 4.5 može se vidjeti kako se signali zvijezda s planetima jasnije razlikuju od onih bez planeta u odnosu na neobrađene signale sa Slike 4.1. Kod svih zvijezda s planetom izraženi su vrhovi u području niskih frekvencija, dok kod zvijezda bez planeta frekventni spektar izgleda vrlo šumovito i podjednako bez istaknutih vrhova. Na ovaj način je primjenom algoritama je pojednostavljena potraga na nekoliko značajki koje se nalaze među niskim frekvencijama, umjesto da moraju prepoznati udubljenja koja se mogu nalaziti na bilo kojem mjestu na grafu.

4.3 Rezultati i analiza

Nakon provjere više algoritama s unaprijed zadanim parametrima logistička regresija i stablo odlučivanja pokazali su se kao dobar odabir za učenje na podacima. Bilo je još nekoliko obećavajućih algoritama koji linearno odvajaju podatke na sličan način kao logistička regresija, no razlog zašto je odabrano stablo odlučivanja je zato što ono može podatke odvajati samo pravcima, odnosno hiper ravninama u višim dimenzijama prostora parametara, koji su vertikalni i okomiti u odnosu na osi prostora



(a)

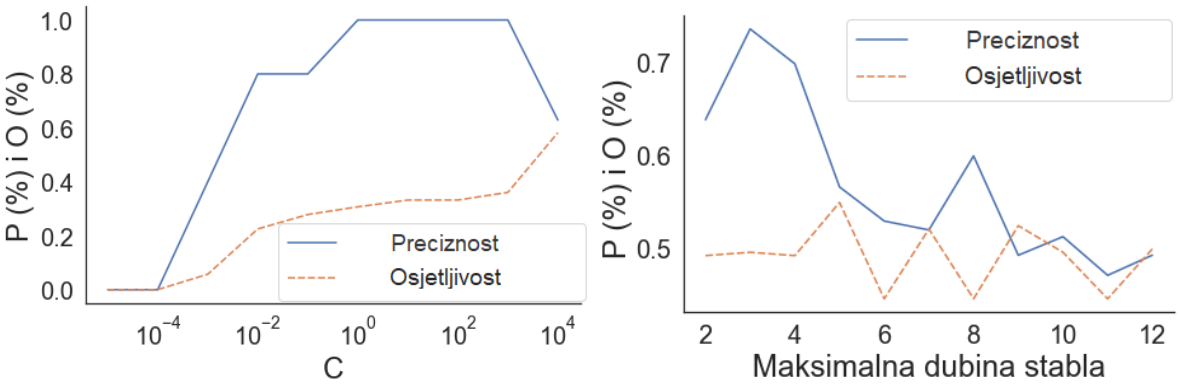


(b)

Slika 4.5: Spektar intenziteta svjetlosti nekih zvijezda s egzoplanetima (a) i bez egzoplaneta (b) nakon obrade podataka.

varijabli što se može vidjeti na Slici 3.11. Hiperravnina logističke regresije može biti orijentirana pod svim kutovima, ali zato odvaja podatke samo jednom hiperravninom za razliku od stabla odlučivanja koje razdvaja podatke sa više njih. To znači da će logistička regresija i stablo odlučivanja raditi različite pogreške te ih je korisno usporediti. Na Slici 4.6 prikazana je ovisnost preciznosti i osjetljivosti algoritama o njihovim parametrima prije balansiranja podataka. Rezultati nisu zadovoljavajući. Htjeli bismo visoke vrijednosti osjetljivosti tijekom unakrsne provjere jer je u skupu za testiranje samo 5 zvijezda s planetom, a 565 zvijezda bez planeta. Zbog toga

nam je bitno imati veliku osjetljivost, dok je preciznost i broj lažnih pozitiva manje važan. U skupu podataka za treniranje je 37 zvijezda s planetom i 5013 bez planeta, te je takav skup podataka jako nebalansiran. To stvara problem algoritmima strojnog učenja i najbolje je imati podjednak broj pozitivno i negativno označenih podataka. U nastavku algoritme učimo na podacima koje balansiramo nasumičnim kopiranjem postojećih redaka zvijezda s planetom u tablicu. Takvih redaka nema jednako mnogo kao redaka sa zvijezdama bez planeta. Ta tehnika zove se tehnika pretjeranog uzorkovanja manjine (engl. Synthetic Minority Over-sampling Technique, SMOTE) [25].



(a) (b)

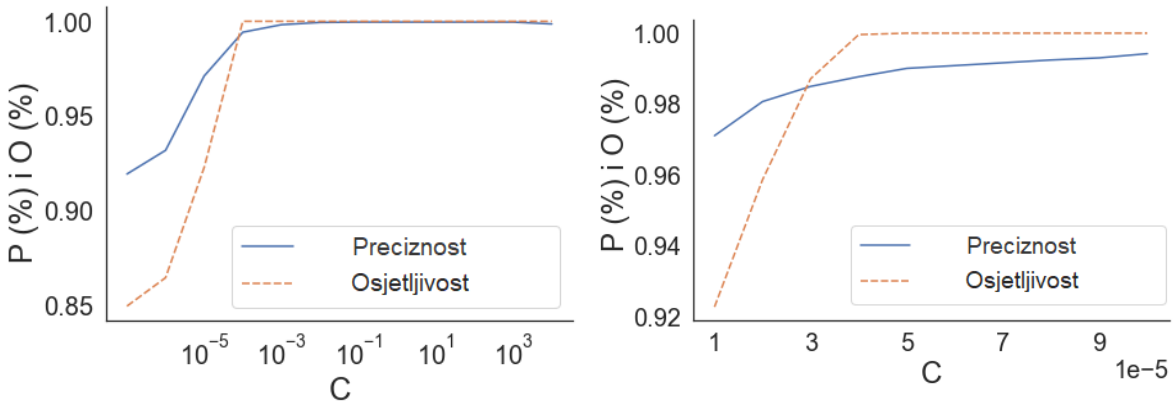
Slika 4.6: Ovisnost preciznosti (P) i osjetljivosti (O) logističke regresije o parametru C (a) i stabla odlučivanja o dubini stabla (b) koristeći nebalansirane podatke. Preciznost i osjetljivost definirane su formulama 3.8 i 3.9.

Rezultati logističke regresije Kako bi pronašli najbolji model strojnog učenja za predviđanje željenih rezultata, nakon obrade podataka i odabira algoritama, potrebno je testirati kako algoritmi ovise o parametrima. Parametar koji proučavamo kod logističke regresije je vrijednost koeficijenta C u regularizacijskom članu funkcije troška $\frac{1}{C}||\mathbf{w}||^2$ pri čemu je $||\mathbf{w}||$ euklidska norma vektora težinskih koeficijenata. Regularizacija je ograničavanje koeficijenata težina algoritma dodavanjem regularizacijskog člana funkciji troška, a cilj ograničavanja je izbjegavanje prenaučenosti algoritma [10]. Funkcija troška logističke regresije u cijelosti je zadana funkcijom:

$$J(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)] + \frac{1}{C}||\mathbf{w}||^2 \tag{4.1}$$

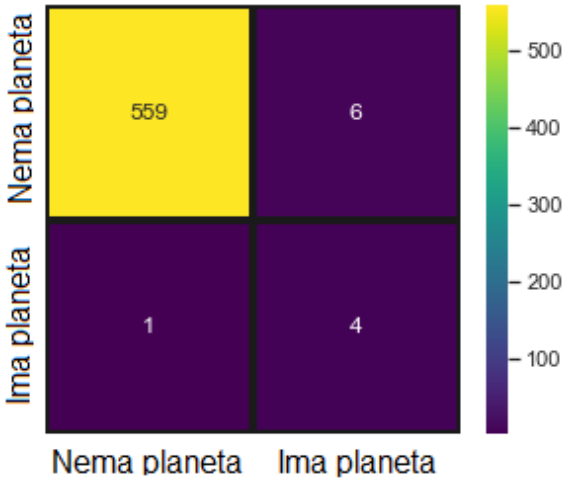
Koristeći manje iznose C regularizacija je veća i zbog toga se vrijednosti težinskih koeficijenata kreću oko nule. Kod većih iznosa C regularizacija je manja i algoritam

ima više slobode u prilagođavanju podacima. Kako bi pronašli najbolji C za logističku regresiju testirali smo preciznost i osjetljivost algoritma na više redova veličina parametra C (Slika 4.7a), a zatim smo detaljnije testirali područje u intervalu $[10^{-5}, 10^{-4}]$ (Slika 4.7b), nakon što je Slika 4.7a pokazala da u tom intervalu preciznost i osjetljivost prestaju rasti. Najbolje je odabrati vrijednost parametra gdje preciznost i osjetljivost prestaju značajno rasti jer se nakon toga može očekivati prenaučenost. Kao najbolji izbor pokazala se vrijednost $C = 4 \cdot 10^{-5}$.



(a) (b)

Slika 4.7: Ovisnost preciznosti (P) i osjetljivosti (O) logističke regresije o parametru C (a) i zumirani interval te ovisnosti (b) za balansirane podatke. Preciznost i osjetljivost definirane su formulama 3.8 i 3.9.



Slika 4.8: Matrica zbunjenosti logističke regresije. Na vertikalnoj osi nalaze se stvarne klase zvijezda i na horizontalnoj osi predviđene klase. Na glavnoj dijagonali matrice (gore lijevo prema dolje desno) predviđene i stvarne vrijednosti se poklapaju, a ostale vrijednosti su krive klasifikacije.

Iz matrice zbunjenosti koja je prikazana na Slici 4.8 može se vidjeti da je logistička regresija točno predvidjela 4 od ukupno 5 zvijezda s egzoplanetom, a uz njih i 6

lažnih pozitiva. U Tablici 4.2 zapisane su osjetljivosti i preciznosti logističke regresije s odabranim parametrom $C = 4 \cdot 10^{-5}$. Logistička regresija ima malu preciznost zbog velikog broja lažnih pozitiva u odnosu na broj stvarnih pozitiva, ali kao što smo već spomenuli u poglavlju 3.2, možemo biti zadovoljni s tim rezultatom jer je jako malo potvrđenih egzoplaneta s obzirom na broj zvijezda. Kad bi ovaj rezultat bio realna situacija, morali bismo ručno provjeriti 10 zvijezda koje je algoritam svrstao kao one s egzoplanetom umjesto svih 570 zvijezda, a za svakih pet predviđenih zvijezda mogli bismo očekivati dvije s egzoplanetom.

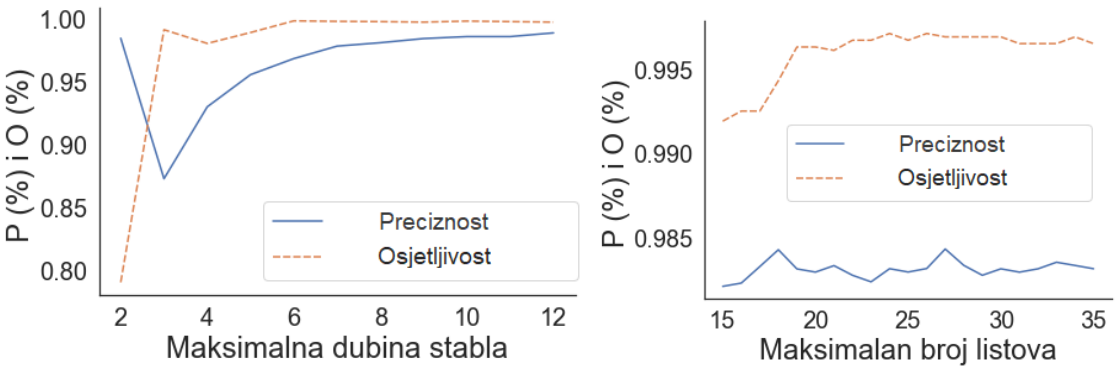
	Osjetljivost	**Preciznost**
Zvijezde s egzoplanetom	0.8	0.4
Zvijezde bez egzoplaneta	0.99	1

Tablica 4.2: Preciznost i osjetljivost logističke regresije.

Rezultati stabla odlučivanja Kao što smo već spomenuli, stabla odlučivanja s neograničenom dubinom i brojem listova postaju prenaučena. Zbog toga su dubina i broj listova glavni parametri koje testiramo. Na Slici 4.9a testira se maksimalna dubina stabla koristeći entropiju kao mjeru nečistoće. Testirali smo i Gini nečistoću kao mjeru. Rezultati se nisu značajno razlikovali te smo odabrali entropiju jer ona teži stvaranju stabla koje je više balansirano [10]. Odabrali smo stablo maksimalne dubine 8 jer u u tom trenutku preciznost prestaje značajno rasti s dubinom, a osjetljivost je već skoro savršena na podacima za provjeru. Stablo dubine 8 je zatim testirano za više vrijednosti maksimalnog broja listova (Slika 4.9b) te je odabrna vrijednost 22 iz istog razloga. Konačni rezultati zapisani su u Tablici 4.3. Stablo odlučivanja ima nešto lošiju osjetljivost i preciznost na zvijezdama s egzoplanetima u usporedbi s logističkom regresijom, ali jednako dobro predviđa zvijezde koje nemaju egzoplanete. Matrica zbunjenosti stabla odlučivanja, s odabranim parametrima za koje su izračunate točnosti iz tablice 4.3, prikazana je na Slici 4.10.

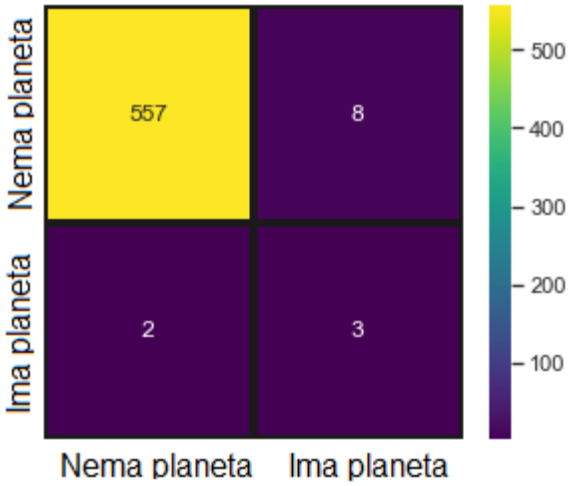
	Osjetljivost	**Preciznost**
Zvijezde s egzoplanetom	0.6	0.27
Zvijezde bez egzoplaneta	0.99	1

Tablica 4.3: Preciznost i osjetljivost stabla odlučivanja.



(a) (b)

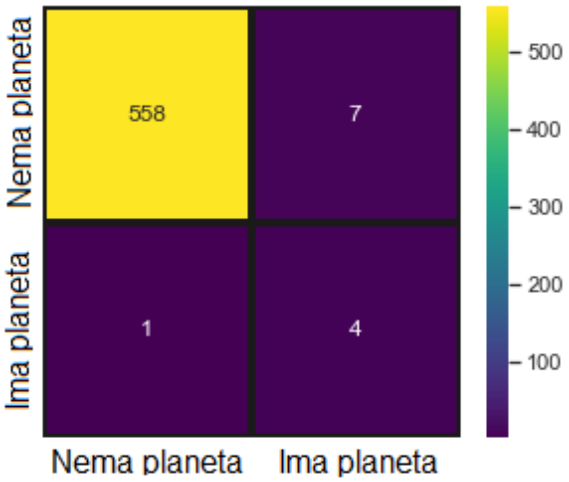
Slika 4.9: Ovisnost preciznosti (P) i osjetljivosti (O) stabla odlučivanja o maksimalnoj dubini stabla (a) i o maksimalnom broju listova (b) koristeći balansirane podatke. Preciznost i osjetljivost definirane su formulama 3.8 i 3.9.



Slika 4.10: Matrica zbunjenosti stabla odlučivanja. Na vertikalnoj osi nalaze se stvarne klase zvijezda i na horizontalnoj osi predviđene klase. Na glavnoj dijagonali matrice (gore lijevo prema dolje desno) predviđene i stvarne vrijednosti se poklapaju, a ostale vrijednosti su krive klasifikacije.

Ansambl logističke regresije i stabla odlučivanja Nešto bolji rezultati mogu se postići korištenjem više algoritama strojnog učenja koji glasuju kako će svrstati podatak prilikom klasifikacije. Općenito, ansambl više algoritama imati će veću preciznost nego svaki algoritam zasebno [10]. Spomenuto je da logistička regresija i stabla odlučivanja rade različite pogreške zbog toga što su im mehanizmi kojim linearno odvajaju podatke sasvim različiti. Usporedbom lažnih pozitiva oba algoritma pokazalo se među lažnim pozitivima podudaraju samo dvije zvijezde, a sve ostale su različite. S obzirom da ansambl ima bolju preciznost kad klasifikatori imaju manje korelirane pogreške, dobra je ideja provjeriti njegove rezultate [10]. Matrica zbunjenosti prikazana je na Slici 4.11, a pripadne točnosti u Tablici 4.4. Ansambl daje

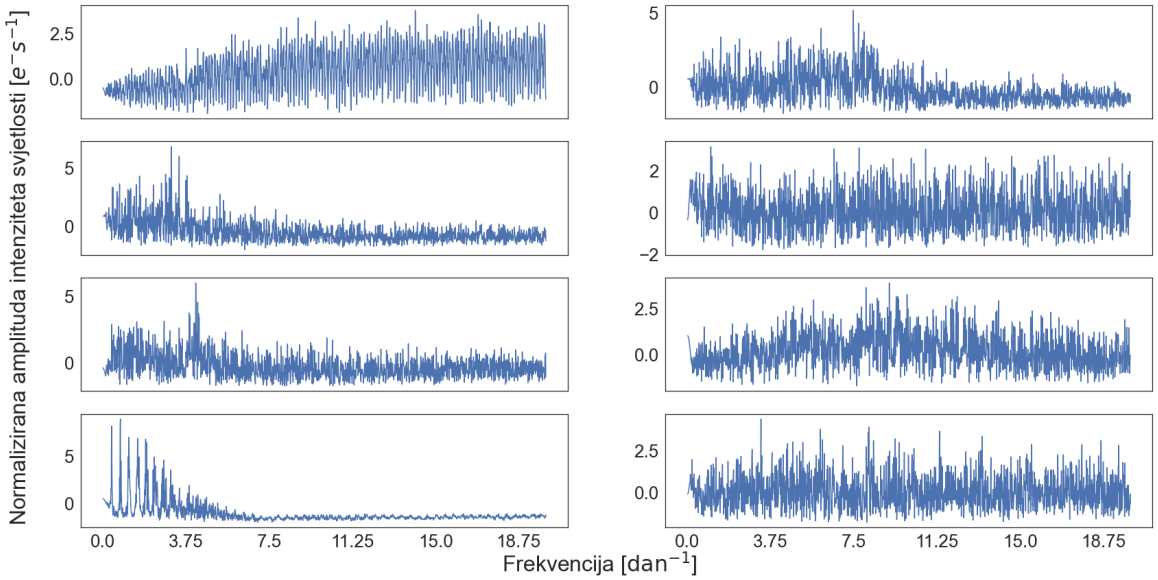
slične rezultate kao logistička regresija, a bolje nego stablo odlučivanja. Lažni pozitivi se razlikuju od oba klasifikatora zasebno i prikazani su na Slici 4.12. Za nekoliko signala frekventni spektar zaista ima vrhove u niskom frekventnom području zbog čega su klasifikatori mogli predvidjeti da su to zvijezde s planetima, a gledajući njihove originalne grafove vidi se kako neki od njih imaju nekoliko uskih vrhova koji će se prikazati u frekventom spektru među niskim frekvencijama na isti način kao da su udubine u grafu jer je uzeta u obzir samo apsolutna vrijednost kod Fourierove transformacije. To su mogući uzroci krivih predviđanja. Takvi rezultati lako se eliminiraju brzim pogledom originalnog signala jer očito nemaju uske udubine u grafu. Zanimljiv rezultat je što se u lažnim pozitivima nalaze dva signala, na Slici 4.12b drugi red desno i zadnji red lijevo, koji imaju karakteristike zvijezda s egzoplanetima. Prvi ima tri manje, široke udubine u grafu dok drugi ima puno uskih udubina koje podsjećaju na neke signale sa Slike 4.1a. Ovo je mogući propust baze podataka jer bi ove zvijezde mogle biti krivo klasificirane kao zvijezde bez planeta. Baza podataka koju koristimo zadnji put je ažurirana prije četiri godine od današnjeg dana (11. 6. 2021.). U sklopu ovog rada nismo u mogućnosti odrediti imaju li zvijezde egzoplanete te bi se to trebalo provjeriti drugim metodama.



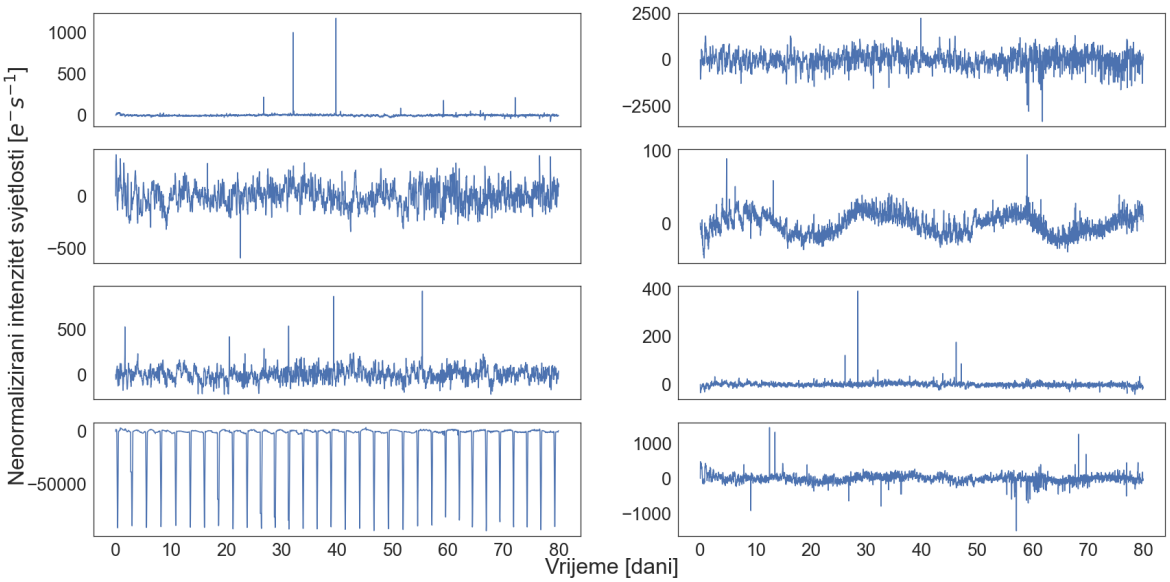
Slika 4.11: Matrica zbunjenosti ansambla logističke regresije i stabla odlučivanja. Na vertikalnoj osi nalaze se stvarne klase zvijezda i na horizontalnoj osi predviđene klase. Na glavnoj dijagonali matrice (gore lijevo prema dolje desno) predviđene i stvarne vrijednosti se poklapaju, a ostale vrijednosti su krive klasifikacije.

	Osjetljivost	Preciznost
Zvijezde s egzoplanetom	0.8	0.36
Zvijezde bez egzoplaneta	0.99	1

Tablica 4.4: Preciznost i osjetljivost ansambla logističke regresije i stabla odlučivanja.



(a)



(b)

Slika 4.12: Obrađeni (a) i neobrađeni (b) signali zvijezda među lažnim pozitivima ansambla.

5 Zaključak

U radu je provedeno istraživanje algoritama strojnog učenja koji bi se mogli primijeniti za klasifikaciju zvijezda s egzoplanetima. Rješenja su napravljena po uzoru na rad jednog korisnika platforme Kaggle čija se obrada podataka nije pokazala kao adekvatna za potrebe strojnog učenja [3]. Originalni podaci i smanjivanje njihove dimenzionalnosti dali su jako loše rezultate na testnom skupu podataka što ih čini jako podnaučenima. Za našu obradu transformirali smo podatke u frekventnu domenu u kojoj se tranzitne pojave pojavljuju kao događaji niskih frekvencija. Na ovaj način algoritmi strojnog učenja nisu zaduženi za prepoznavanje udubina u krivulji intenziteta svjetlosti koje se mogu pojaviti bilo gdje na krivulji, već procjenjuju na temelju toga postoje li izražene niske frekvencije u frekventnom spektru signala. Zbog velike nebalansiranosti u skupu podataka koristili smo tehniku pretjeranog uzorkovanja manjine kako bismo umjetno stvorili velik broj podataka s potvrđenim egzoplanetima.

Kao najbolji algoritmi za naš slučaj pokazali su se logistička regresija i stablo odlučivanja koji su postigli zadovoljavajuće rezultate. Logistička regresija točno je klasificirala četiri od pet egzoplaneta u skupu podataka za testiranje, a stablo odlučivanja tri od pet. Oba algoritma predviđaju otprilike dvostruko više lažnih pozitiva od stvarnog broja egzoplaneta, ali u astronomiji je to prihvatljivo. S obzirom na manjak zvijezda s egzoplanetima u skupu podataka za treniranje i testiranje, traženje egzoplaneta je traženje igle u plastu sijena. Pošto su algoritmi uspjeli naći nekoliko egzoplaneta, smatramo klasifikaciju vrlo uspješnom.

Za bolje rezultate predlažemo korištenje ansambla logističke regresije i stabla odlučivanja, te ako je moguće još nekoliko algoritama strojnog učenja kako bi se povećala preciznost klasifikacije. S većom količinom podataka također bi se mogli postići bolji rezultati, posebice s većim brojem potvrđenih egzoplaneta kojih nedostaje u skupu podataka. Kao zadnje moguće unaprijeđenje predlažemo pojedinačnu obradu podataka jer je za ovaj skup podataka vrlo teško odabrati zajedničku transformaciju koja garantira da će rezultati klasifikacije općenito biti dobri. Zbog ovih problema znanstvenici već pribjegavaju opciji znanosti za građanstvo koja je opisana u šestom poglavlju. Zajednički rad velikog broja ljudi je u nekim slučajevima efikasniji od automatiziranih algoritama.

6 Metodički dio

6.1 Projekti znanosti za građanstvo

Znanost za građanstvo (engl. citizen science) je zajednički pojam za znanstvena istraživanja u kojima sudjeluju amateri. Sve aktivnosti organiziraju i koordiniraju znanstvene institucije koje provode određena istraživanja. Takvi projekti od velikog su značaja za popularizaciju znanosti jer se kroz njih educiraju ljudi i povećava se razumijevanje znanosti u društvu [26]. Projektima znanosti za građanstvo mogu pristupiti ljudi svih uzrasta, s različitim obrazovanjem i znanjem. To je izvrsna prilika za uključivanje učenika i mladih u znanstvene projekte. Mladi takvim radom razvijaju zanimanje za znanost i mogu izabrati određeno područje kao svoje buduće zanimanje.

Najviše projekata znanosti za građanstvo je iz područja prirodnih znanosti, kao što su fizika i biologija, te iz računalnih znanosti i matematike. Veći broj projekata je iz astronomije, najstarije znanosti koja fascinira ljude od davnih dana dok još pojam znanstvenika nije ni postojao. Astronomija je odličan primjer utjecaja amatera koji često promatraju i otkrivaju nove objekte na nebu.

Specijalna vrsta projekata znanosti za građanstvo su projekti distribuiranog računarstva [27]. S porastom primjene modernih računalnih metoda, kao što su računarstvo visokih performansi i strojno učenje, potrebno je obraditi sve veću količinu podataka što nije jednostavan ni jeftin zadatak. Potrebu za velikom računalnom snagom u znanosti je teško zadovoljiti. Projekti znanosti za građanstvo unaprjeđuju znanost na skali koja nije bila moguća prije izuma interneta. U projektima distribuiranog računarstva aktivni članovi prikupljaju, analiziraju, ili prepisuju podatke, a sve što im je potrebno je pristup internetu i osobnom računalu. Nekim projektima može se pomoći prikupljanjem podataka pomoću aplikacije na pametnim telefonima.

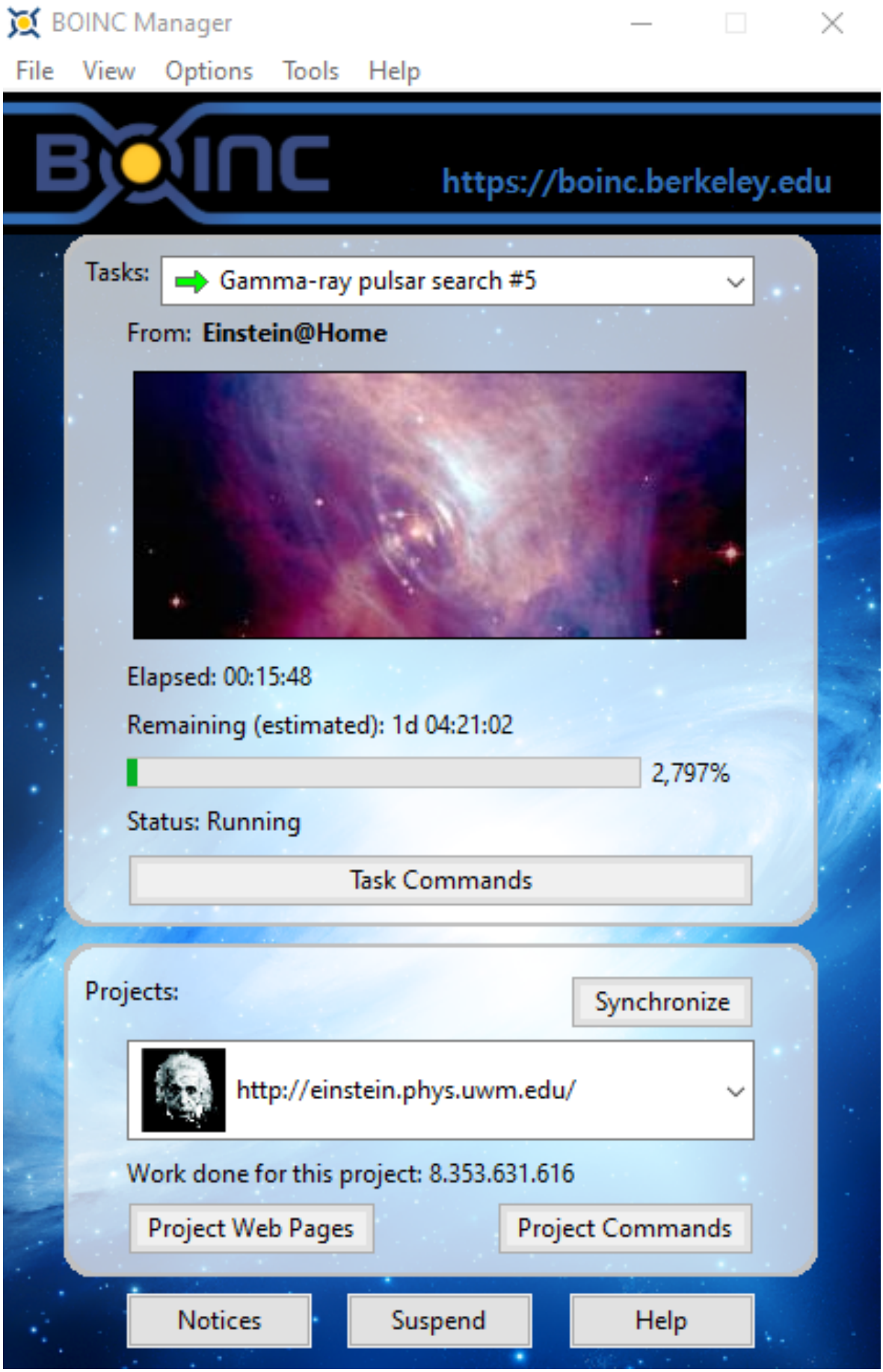
U sustavu distribuiranog računarstva sudionici dobrovoljno doniraju procesorsko vrijeme svojih osobnih računala. Takvi sustavi sastoje se od središnjeg programa koji korisnici instaliraju na svoja računala i servera nadležne institucije koji su zaduženi za podjelu zadataka svim umreženim računalima [27]. Najbitnija karakteristika distribuiranog računarstva je da se donira samo procesorsko vrijeme praznog hoda (engl. idle time), odnosno ono procesorsko vrijeme koje inače ne bi bilo iskorišteno. Kada ne bi tako bilo, nitko ne bi ni htio sudjelovati u takvim projektima. Zanimljiva poslje-

dica je da se na ovaj način iskorištava energija koja inače ne bi bila iskorištena. Za one koji ostavljaju računala uključena cijeli dan, iako ih ne koriste stalno, ovo je izvrsna prilika za doprinos znanstvenoj zajednici. Sudionici se potiču praćenjem poklonjenog procesorskog vremena sustavom bodova i proglašavanjem sudionika mjeseca, te pristupom svim rezultatima projekta koje su pomogli ostvariti.

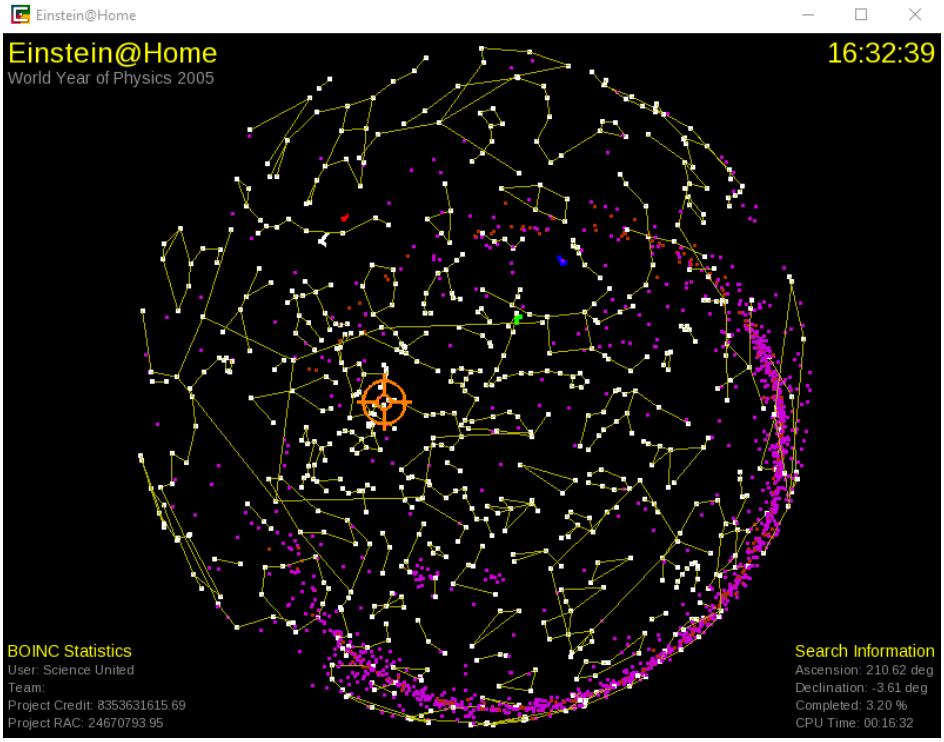
Prvi projekt distribuiranog računarstva iz područja fizike bio je SETI@home [28, 29]. Cilj ovog projekta je bio potraga za znakovima inteligentnog izvanzemaljskog života. Analizirali su se radio signali snimani pomoću Arecibo i Green Bank teleskopa [29]. Jedan od važnijih programa distribuiranog računarstva, koji omogućuje pristup mnoštvu projekata, zove se BOINC (engl. Berkeley Open Infrastructure for Network Computing) [30]. Na Slici 6.1 prikazano je korisničko sučelje programa BOINC na kojemu je potrebno u padajućem izborniku odabrati projekt kojemu želimo pokloniti procesorsko vrijeme nakon čega će se automatski dodijeliti zadaci računalu dok ga vlasnik ne koristi. Poznata je i platforma projekata distribuiranog računarstva Zooniverse koja u ovom trenutku (30. 6. 2021.) ima 2325818 registriranih članova [31]. Neki projekti distribuiranog računarstva podržavaju vizualizaciju podataka koje obrađujemo, kao što je to slučaj kod Einstein@home projekta (Slika 6.2) [32]. U tom projektu se traže slabi signali pulsara koristeći podatke LIGO (The Laser Interferometer Gravitational-Wave Observatory) detektora gravitacijskih valova. Na stranici projekta može se pristupiti popisu otkrivenih pulsara i njihovim rezultatima pored kojih pišu imena osoba koje su pomogle to ostvariti na svojim računalima [33]. Neki od najuspješnijih projekata na platformi BOINC su Rosetta@home (predviđanje strukture proteina za istraživanje bolesti), Asteroids@home (istraživanje svojstava asteroida), i MilkyWay@home (stvaranje trodimenzionalnog modela naše galaksije iz podataka Sloanovog pregleda neba) [28].

6.2 O projektu Planet Hunters TESS

Transiting Exoplanet Survey Satellite (TESS) Misija TESS koju vodi MIT u suradnji s NASA-om dobila je naziv po istoimenom svemirskom teleskopu u Zemljinoj orbiti [4]. Teleskop snima zvjezdano nebo koje je podijeljeno u 26 segmenata, a svaki segment snima se različit broj dana zbog ograničenja koje predstavlja orbita teleskopa. Većina segmenata snimana je ukupno 27 dana, što teleskop čini najosjet-



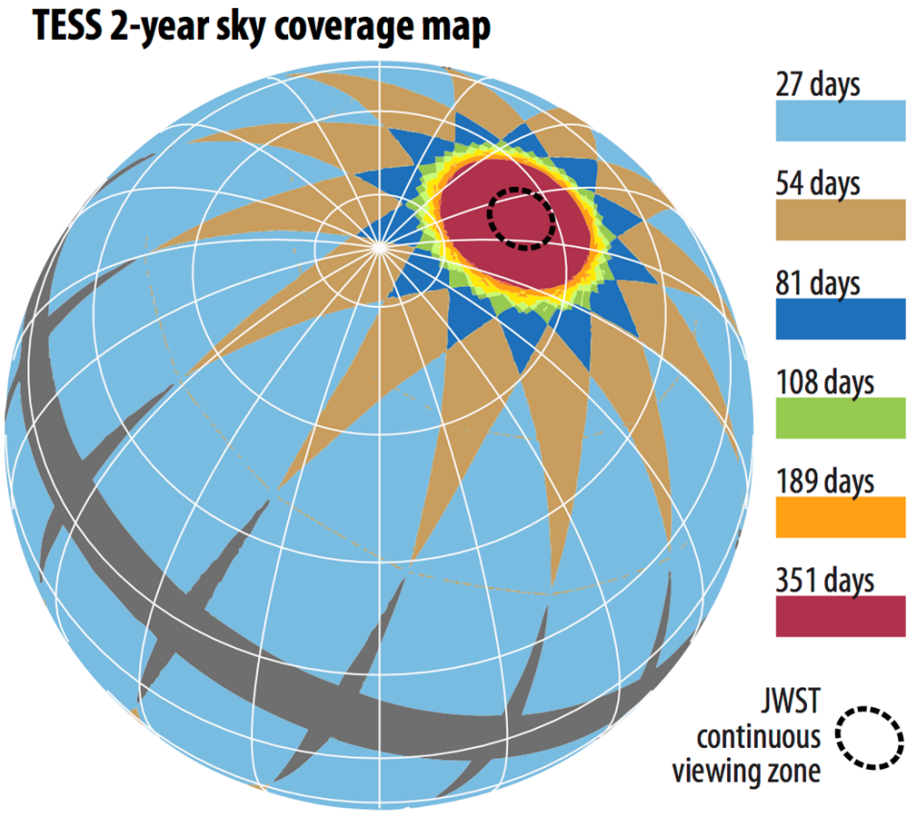
Slika 6.1: Korisničko sučelje programa BOINC [30].



Slika 6.2: Vizualizacija podataka jednog od zadataka Einstein@home projekta [32].

ljivijim na egzoplanete perioda manjeg od 13 dana kod kojih bi se tranzit dogodio barem dva puta za vrijeme snimanja. Kod segmenata koji se promatraju duže vrijeme povećava se vjerojatnost pronalaska egzoplaneta većih perioda orbite. Kompletna podjela zvjezdanog neba na segmente prikazana je na Slici 6.3. Satelit snima s četiri identične kamere visoke rezolucije koje su osjetljive na svjetlost između 600 nm i 1000 nm. U svakom sektoru snima se oko 200000 slika. Nakon uspjeha Kepler misije, očekivalo se da će projekt TESS pronaći tisuće novih egzoplaneta. Glavna misija TESS-a završila je 4. 7. 2020., ali je projekt još uvijek aktivan. Rezultat TESS-a do dašanjeg dana (1. 7. 2021.) je 131 potvrđen egzoplanet i još 4195 kandidata koji se moraju potvrditi dodatnom analizom [34].

Planet Hunters TESS Zbog ograničenog vremena snimanja segmenata automatizirani sustavi za detekciju planeta dugih perioda ne funkcioniraju dobro. Ti sustavi najbolje funkcioniraju kada postoji nekoliko tranzita tijekom snimanja, te im problem predstavljaju slučajevi kada je snimljen samo jedan tranzit egzoplaneta. Alternativne metode, kao što su strojno učenje ili znanost za građanstvo, u nekim slučajevima mogu nadmašiti standardne metode otkrivanja planeta. Konkretno, ljudi su učinkovitiji u pronalaženju pojedinačnih tranzita egzoplaneta nego algoritmi za

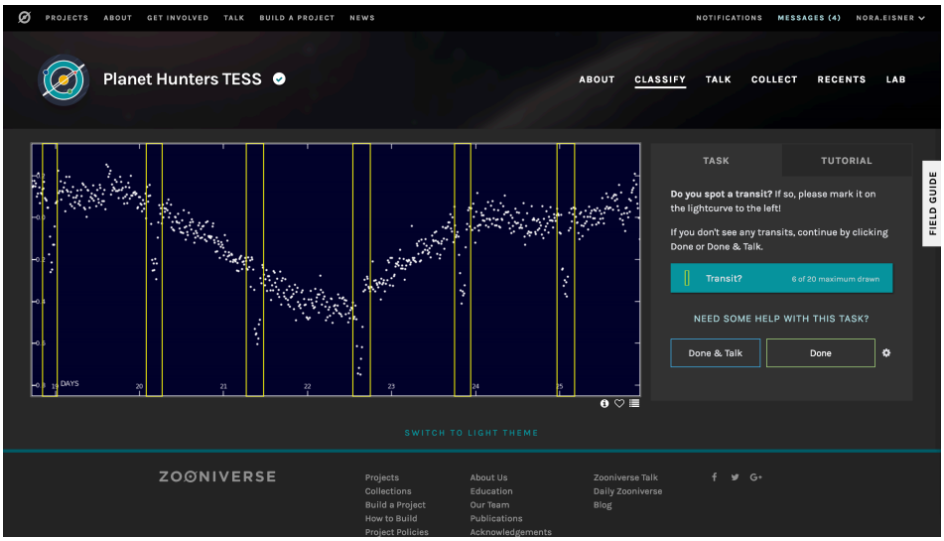


Slika 6.3: Prikaz predviđenih segmenata neba koje snima TESS [4].

automatsku detekciju [35, 36].

Taj problem motivirao je otvaranje projekta Planet Hunters TESS [5]. Glavni cilj projekta je angažiranje velikog broja amatera koji suradnjom pronalaze tranzite egzoplaneta koje standardni sustav za detekciju nije uspio otkriti. Projekt ima svoju web stranicu na kojoj se korisnici mogu prijaviti, ili sudjelovati anonimno. Na njoj se nalaze detaljne upute o projektu i softverskim alatima koji olakšavaju rad. Osnovno sučelje koje korisnici koriste prilikom rada (Slika 6.4) sastoji se od prozora s prikazom krivulje intenziteta svjetlosti na kojoj korisnici označavaju tranzitne pojave koje se pojavljuju kao udubljenja u krivulji intenziteta svjetlosti zvijezde. Pored toga postoji sustav za komunikaciju s drugim korisnicima i znanstvenicima kojem se može pristupiti klikom na gumb. Projekt Planet Hunters TESS omogućio je da svi koji to žele na svoje računalo snime kompletne podatke o intenzitetima svjetlosti i samoinicijativno rade detaljnu analizu [35].

Planet Hunters TESS kao projekt popularizacije znanosti u društvu Uspjeh projekta Planet Hunters TESS proizlazi iz nekoliko zanimljivih faktora koji su potaknuli sudjelovanje. Pokazalo se da je za sve projekte na platformi Zooniverse motivacija



Slika 6.4: Web sučelje Planet Hunters TESS projekta [35].

korisnika usko vezana sa željom da ostvare doprinos znanosti, odnosno da ostvare određeno postignuće u sudjelovanju koje mogu prepoznati drugi. Nije nužno da se ljudi jako zanimaju za područje kojem pomažu, već motivacija vjerojatno proizlazi iz želje za sudjelovanjem [37]. No ako je sudjelovanje kompliciranо, ono neće privući puno korisnika. Planet Hunters TESS uspio je ostvariti velik interes motiviranih i sposobnih ljudi koji su postigli kvalitetne rezultate. Jedan od razloga je slobodan pristup podacima koji je omogućen svim korisnicima na platformi. Neki astronomski projekti distribuiranog računarstva, kao što je Supernova Hunters [38], nisu otvorili pristup svim podacima kako bi korisnici mogli samostalno raditi s njima te nisu razvili tako naprednu i motiviranu zajednicu. Razvojem softverskih alata, u koji je projekt Planet Hunters TESS uložio vrijeme i novac, olakšava se pristup većem broju korisnika što ima jednaku važnost kao i pristup podacima. Jednostavan pristup alatima i podacima bio je dakle ključan za uspjeh projekta. Sudionici su za dva tjedna nakon objave podataka klasificirali sve zvijezde svakog pojedinog sektora [35]! Ovakav pristup znanosti za građanstvo očito jako dobro funkcionira te je izvrsna prilika za poticanje interesa mladih za znanost i astronomiju. Jedan od sudionika, Cesar Rubio strojar u SAD, sudjelovao je u projektu sa svojim sedmogodišnjim sinom Miguelom koji s tatom voli pričati o planetima i zvijezdama [39]. Sudjelovali su u raspravi i zajedno pomogli klasificirati planetarni sustav zvijezde s dva egzoplaneta zbog čega su bili koautori članka u kojem je opisano to otkriće. To je iskustvo koje sedmogodišnji Miguel neće nikad zaboraviti i koje mu nitko ne može kupiti. Jednog dana ovo će možda biti prekretnica u Miguelovom odabiru karijere. Ovakvi projekti imaju

značajan utjecaj na mlade i popularizaciju znanosti.

Projekt Planet Hunters TESS u školama S obzirom da je Planet Hunters TESS osmišljen tako da se sučelje jednostavno koristi, projekt je dobra opcija za fakultativnu nastavu u školi. Nastavnik bi prvo održao kraće predavanje o egzoplanetima i potragom za životom izvan Zemlje. Uz pomoć nastavnika učenici bi brzo naučili kako se koristiti web sučelje i što je njihov zadatak. Cilj je potaknuti u učenicima želju za istraživanjem i pojasniti im značaj projekta. Podatke intenziteta svjetlosti zvijezda na kojima učenici rade moguće je snimiti na osobno računalo i samostalno proučiti vlastitom obradom podataka. Obrada nije jednostavna za učenike, te bi bilo potrebno pripremiti unaprijed programske isječke kojima se transformiraju podaci. Učenicima se može pokazati kako astronomi iz krivulje mogu procijeniti period orbite planeta, ili njegove dimenzije. To bi se moglo postići kroz razrednu raspravu kako bi se učenike potaknulo na razmišljanje i pitanja. Moguća pitanja su:

- Što znači ako se u krivulji nalazi više tranzitnih pojava, ali su različitih dubina? To bi mogao biti znak da postoji više planeta.
- Što bi još mogao biti znak da zvijezda ima više od jednog planeta? Ako tranzitne pojave nisu jednoliko razmaknute postoji mogućnost da je više planeta oko zvijezde.
- Kakve periode orbita imaju planeti?

Projekt je vrlo zgodan za povezivanje školskih predmeta i razvija interdisciplinarnost. Prilikom sudjelovanja učenici bi razvijali vještine iz fizike, astronomije, biologije i informatike. Razvijale bi se i vještine engleskog jezika i znanstvene komunikacije jer s novim saznanjima pripremamo učenike za raspravu na stranicama Planet Hunters TESS projekta koja je uspješan faktor pri pronalaženju zanimljivih događaja na nebu. Opcija za raspravu, zvana Talk u web aplikaciji, zaslužna je za pronalazak neobičnih objekata koji se nisu uspjeli klasificirati automatiziranim metodama. Zbog aktivnosti u raspravi privučena je dodatna pažnja na te događaje. Tijekom projekta otkrića su često započela raspravom unutar Talk opcije, a ne samo analizom označenih podataka [37].

Ovom fakultativnom nastavom želimo potaknuti interes i razviti istraživačke vještine učenika. Uzbudljivo je to što je Planet Hunters TESS pravi znanstveni projekt koji

doprinosi znanju čovječanstva. Učenici se ne bi ocjenjivali, osim poticajnih petica za sudjelovanje u fakultativnoj nastavi. Ako se otkriće dogodi baš u vašem razredu, s kolegama ćete dijeliti jedinstveno iskustvo koje nikad nećete zaboraviti.

6.3 Nastavna priprema: Newtonov zakon gravitacije

ŠKOLA: Srednja škola, opća gimnazija
RAZRED: 1. razred
NASTAVNA JEDINICA: Opći zakon gravitacije
PREDVIĐENI BROJ SATI: 2
PREDMETNI ISHODI:
FIZ SŠ C.1.7. Primjenjuje zakon gravitacije i analizira gibanje Zemlje i nebeskih tijela.

- Navodi Newtonov zakon gravitacije i objašnjava njegovo značenje.
- Navodi u kojim situacijama vrijedi Newtonov zakon te koje su pretpostavke modela.
- Opisuje razlike u težini tijela na različitim planetima Sunčevog sustava.
- Primjenjuje Newtonov zakon gravitacije

MEĐUPREDMETNI ISHODI:
osr B.4.2

- Suradnički uči i radi u timu.

uku B.4/5.1.

- Planiranje: učenik samostalno određuje ciljeve učenja, odabire pristup učenju te planira učenje.

B.4.1.A

- Odabire primjerene odnose i komunikaciju.

ikt B.4.3.

- Učenik kritički procjenjuje svoje ponašanje i ponašanje drugih u digitalnome okružju.

pod B.4.2.

- Planira i upravlja aktivnostima.

B.4.1.A

- Odabire primjerene odnose i komunikaciju.

VRSTA NASTAVE: Istraživački usmjerena nastava
NASTAVNE METODE:

- Učeničko izvođenje pokusa /mjerenja u skupinama
- Metoda razgovora - razredna rasprava
- Konceptualna pitanja s karticama
- Metoda pisanja /crtanja

OBLICI RADA:

- Frontalni
- Individualni
- Rad u skupinama

KORELACIJA S DRUGIM PREDMETIMA: Matematika.
NASTAVNA POMAGALA I SREDSTVA: Računalo, projektor, tableti, ploča.
LITERATURA: Literatura korištena za izradu ove pripreme navedena je u bibliografiji [40–43]. Za izvođenje pokusa koristi se simulacija [44].

TIJEK NASTAVNOG SATA

Uvodni dio sata Uvodni problem: Kolikom silom Zemlja privlači Mjesec? Ako je masa Mjeseca $m = 7,347 \cdot 10^{22}$**,** kolikom silom Zemlja privlači Mjesec?

$$F_g = mg = 7,347 \cdot 10^{22}\ \text{kg} \cdot 9,81\ \frac{\text{N}}{\text{kg}} = 7,21 \cdot 10^{23}\ \text{N}$$

Usporedite dobivenu vrijednost s tabličnom vrijednosti $F = 1,98 \cdot 10^{20}$ N.

U kakvom su odnosu vrijednost sile koju ste izračunali i tablična vrijednost?

U kojim slučajevima se može koristiti formula za silu težu? Može li se primijeniti za računanje Zemljine sile na Mjesec?

Što sugerira manja vrijednost gravitacijske sile kod tablične vrijednosti?

Razrednom raspravom učenike se potiče na razmišljanje prethodnim pitanjima o uvodnom problemu nakon čega dolaze do nekoliko zaključaka:

- Koristeći silu težu dobio se oko tisuću puta veći iznos sile od stvarnog iznosa.
- Korištena formula za silu težu primjenjuje se kod računanja gravitacijske sile na predmete blizu površine Zemlje, a Mjesec se nalazi na velikoj udaljenosti.
- Koristeći izraz za silu težu pretpostavili smo da se Mjesec nalazi na površini Zemlje što znači da manji iznos sile, uzevši u obzir veliku udaljenost Mjeseca od Zemlje, sugerira da je gravitacijska sila manja na većoj udaljenosti.
- Da bismo izračunali gravitacijsku silu na Mjesec moramo saznati kako gravitacijska sila djeluje na većim udaljenostima.

Uvodi se naslov nove nastavne jedinice koja će se proučavati na satu: Newtonov zakon gravitacije.

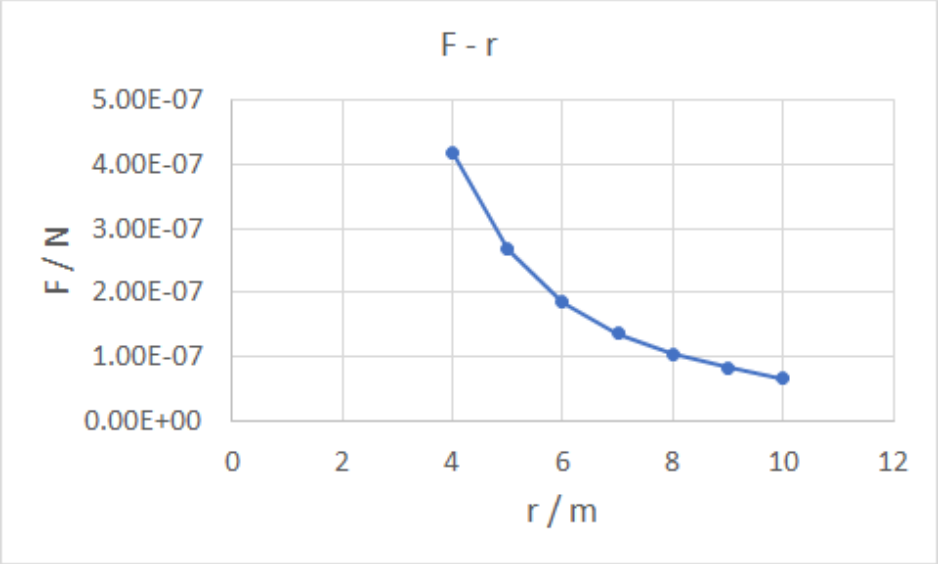
Središnji dio sata Istraživačka pitanja u ovom dijelu sata provode se pomoću simulacije koja je slobodno dostupna na webu [44]. Ovisno o mogućnostima u školi, pokusi se mogu izvoditi frontalno koristeći projektor za prikaz simulacije, ili u grupama. Ako je moguće, podijeliti razred na grupe po pet učenika i svakoj grupi dati tablet na kojemu će moći otvoriti simulaciju i provoditi pokuse. Kod grupnog rada obilaziti učenike i poticati pitanjima, a kod frontalnog rada provoditi pokus zajedno s učenicima kroz razrednu raspravu. Mogu se zapisati mjerenja nekoliko grupa u Excel tablicu i nacrtati dijagrame kako bi pomoću projektora usporedili rezultate grupa.

IP1: Kako gravitacijska sila ovisi o udaljenosti između dvaju tijela? Učenici opisuju i analiziraju postav eksperimenta. Prepoznaju dva tijela koja se nalaze na nekoj udaljenosti, te da djeluju gravitacijskom silom jedno na drugo. Prepoznaju relevantne veličine, a potom vrše kontrolu varijabli tako što odaberu konstantne mase oba tijela. Mjere minimalno sedam iznosa sila F_{12} tijela mase m_1 na tijelo mase m_2 za različite udaljenosti r te pišu podatke u tablicu i crtaju F - r dijagram.

Koliki je iznos sile tijela mase m_2 na tijelo mase m_1? Obrazložite.

r / m	F_{12} / N
10	$6,67 \cdot 10^{-8}$
9	$8,24 \cdot 10^{-8}$
8	$1,04 \cdot 10^{-7}$
7	$1,36 \cdot 10^{-7}$
6	$1,85 \cdot 10^{-7}$
5	$2,67 \cdot 10^{-7}$
4	$4,17 \cdot 10^{-7}$

Tablica 6.1: Primjer tablice s mjerenjima za IP1.



Slika 6.5: Prikaz podataka iz Tablice 6.1.

Kako gravitacijska sila ovisi o udaljenosti tijela?

Što mislite, kako bismo matematički mogli opisati ovisnost?

Što bi se dogodilo sa iznosom sile kada bi tijela bila jako udaljena? Bi li sila ikad postala nula?

Što bi se dogodilo sa iznosom sile kada bi udaljenost između tijela bila nula? Je li to realna situacija? Može li sila biti takvog iznosa?

Razrednom raspravom učenike se potiče na razmišljanje nakon čega bi se trebala oformiti objašnjenja:

- Prema trećem Newtonovom zakonu iznosi sila su jednaki $F_{12} = F_{21}$, ali sile djeluju u suprotnim smjerovima.
- Gravitacijska sila između dva tijela pada s udaljenosti između tijela.
- Gravitacijska sila je inverzno proporcionalna s kvadratom udaljenosti: $F \sim \frac{1}{r^2}$.

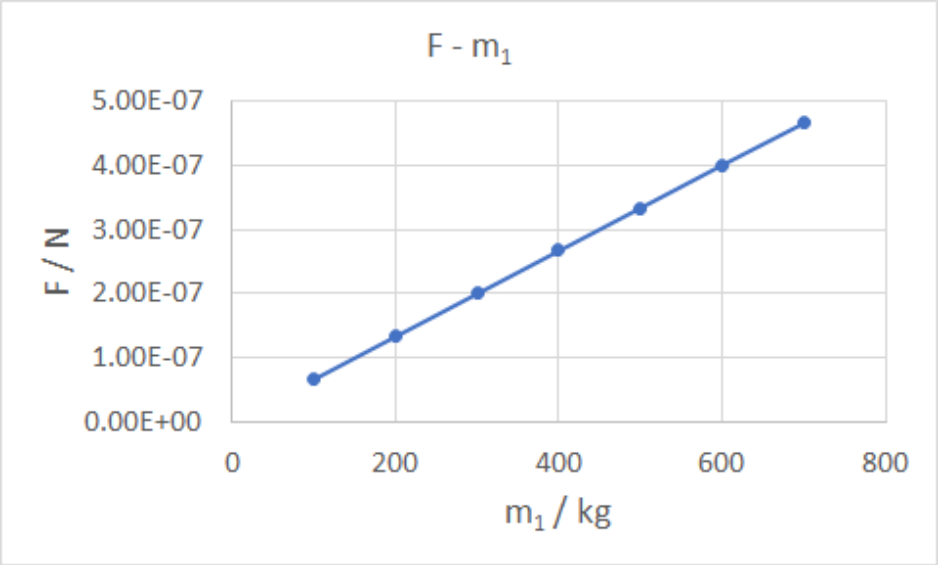
- Sila nikada ne postane nula osim u beskonačnosti. Gravitacijska interakcija je dugodosežna sila. Ta sila djeluje između svih tijela i posvuda je, ali je na velikim udaljenostima ne zamjećujemo zbog iznimno malih iznosa.

- Iznos sile postao bi beskonačan, kada bi udaljenost između tijela bila nula. Beskonačnu silu nije moguće postići i nije moguće staviti dva tijela u potpuno isti položaj u prostoru.

IP2: Kako gravitacijska sila ovisi o masi jednog od tijela? Postav eksperimenta ostaje isti kao u prethodnom pokusu, ali sada drže udaljenost između tijela konstantnom i masu drugog tijela m_2. Mjere gravitacijsku silu prvog tijela na drugo tijelo F_{12} za minimalno sedam različitih masa prvog tijela m_1.

m_1 / kg	F_{12} / N
100	$6,67 \cdot 10^{-8}$
200	$1,34 \cdot 10^{-7}$
300	$2 \cdot 10^{-7}$
400	$2,67 \cdot 10^{-7}$
500	$3,34 \cdot 10^{-7}$
600	$4 \cdot 10^{-7}$
700	$4,67 \cdot 10^{-7}$

Tablica 6.2: Primjer tablice s mjerenjima za IP2.



Slika 6.6: Prikaz podataka iz Tablice 6.2.

Kako gravitacijska sila ovisi o masi tijela mase m_1?

Kako možemo matematički opisati ovu ovisnost?

Što bi se dogodilo kad bismo mijenjali masu drugog tijela m_2? Isprobajte u simulaciji.

Kakav je iznos sile za tijela vrlo male mase?

Razrednom raspravom učenike se potiče na razmišljanje nakon čega bi se trebala oformiti objašnjenja:

- S porastom mase m_1 raste gravitacijska sila kojom tijela međudjeluju.
- Oblik dijagrama je pravac. Gravitacijska sila je dakle proporcionalna s masom tijela m_1: $F \sim m_1$.
- Svejedno je kojem tijelu mijenjamo masu, iznosi sila se identično mijenjaju. Stoga također vrijedi da je sila proporcionalna s masom drugog tijela: $F \sim m_2$.
- Tijela vrlo male mase djeluju vrlo slabom gravitacijskom silom. Za tijela malih masa gravitacijsko djelovanje je neprimjetno.

Pomoću pronađenih proporcionalnosti kako bismo matematički mogli zapisati iznos gravitacijske sile između neka dva točkasta tijela?

Ako se dva točkasta tijela mase m_1 i m_2 nalaze na međusobnoj udaljenosti r, između njih djeluje privlačna gravitacijska sila određena izrazom:

$$F = G\frac{m_1 m_2}{r^2},$$

pri čemu je $G = 6,67 \cdot 10^{-11}\ \frac{\mathrm{Nm}^2}{\mathrm{kg}^2}$ univerzalna gravitacijska konstanta. Taj izraz naziva se Newtonov zakon gravitacije.

Što nam ovaj izraz zapravo govori?

U kojim situacijama vrijedi ovaj zakon prema našim istraživanjima? Možemo li sada izračunati silu Zemlje na Mjesec iz uvodnog problema?

Učenike se potiče još jednom na razrednu raspravu o novim saznanjima iz istraživačkih pokusa i vraća se na uvodni problem s novim znanjem o djelovanju gravitacije.

- Gravitacijska privlačna sila između dvaju tijela ovisi o masi oba tijela i njihovoj udaljenosti. Ta sila je veća što su tijela masivnija, a smanjuje se s kvadratom njihove udaljenosti.

- Newtonov zakon gravitacije vrijedi za tijela kojima je udaljenost velika u usporedbi s njihovim dimenzijama. Odnosno, zakon vrijedi za točkasta tijela, a ako je međusobna udaljenost između neka dva tijela puno puta veća od njihove veličine ona su približno kao točkasta tijela.

- Koristeći Newtonov zakon gravitacije može se izračunati sila Zemlje na mjesec koja odgovara tabličnoj vrijednosti:

$$F = G\frac{m_Z m_M}{r^2} = 6,67428{\cdot}10^{-11}\,\frac{\mathrm{Nm}^2}{\mathrm{kg}^2}\frac{5,972\cdot 10^{24}\cdot 7,347\cdot 10^{22}\ \mathrm{kg}^2}{384401000^2\ \mathrm{m}^2} = 1,98{\cdot}10^{20}\ \mathrm{N}$$

Od čega se sastoje tijela? Je li svejedno ako su oblika kugle, ili su točkasta?
Kako djeluje svaka čestica jedne kugle na drugu kuglu?
Kako bi se mogao dobiti ukupan efekt jedne kugle na drugu?

- Svaka čestica jedne kugle gravitacijskom silom privlači svaku česticu druge kugle. Također vrijedi obratno.

- Zbrajanjem svih pojedinih sila čestica kugle potrebno je izračunati rezultantnu silu. Kada se to napravi ispada da kugla jednolike gustoće privlači svako točkasto tijelo izvan kugle takvom silom kao da je ukupna masa kugle koncentrirana u njezinu središu.

- Sila između dva kuglasta tijela je ista kao da je masa jedne i druge kugle koncentrirana u njihovim središtima.

- Zato se primjenjuje zakon gravitacije koji vrijedi za gravitacijsku silu između točkastih tijela.

Završni dio sata Prva tri zadatka izvode se kao konceptualna pitanja s karticama prilikom kojih se prozove učenike da objasne i pokažu svoja razmišljanja na ploči. Četvrti zadatak zgodan je za raspravu i povezivanje s prethodnim gradivom o kružnom gibanju, a peti je kompliciraniji računski primjer.

1. Dva tijela jednakih masa međusobno su udaljena za r. Udvostručimo li njihove mase, gravitacijska sila među njima bit će

   a) jednaka kao i prije.

b) dvostruko veća nego prije.

c) **četiri puta veća.**

d) četiri puta manja.

$$F_1 = G\frac{m_1 m_2}{r^2}$$
$$F_2 = G\frac{2m_1 2m_2}{r^2} = G\frac{4m_1 m_2}{r^2}$$
$$F_2 = 4F_1$$

2. Dva tijela nepoznatih masa međusobno su udaljena za r. Povećamo li tu udaljenost tri puta, gravitacijska sila između tijela bit će

a) tri puta manja.

b) tri puta veća nego prije.

c) **devet puta manja.**

d) jednaka kao i prije.

$$F_1 = G\frac{m_1 m_2}{r^2}$$
$$F_2 = G\frac{m_1 m_2}{(3r)^2} = G\frac{m_1 m_2}{9r^2}$$
$$F_2 = \frac{1}{9}F_1$$

3. Promotrimo dva tijela jednakih masa na međusobnoj udaljenosti r. Gravitacijska sila između tijela ostaje nepromijenjena ako:

a) masu svakog tijela povećamo 2 puta, a udaljenost 4 puta;

b) udvostručimo udaljenost, ali i masu jednog tijela;

c) smanjimo udaljenost onoliko puta koliko povećamo masu tijela;

d) **masu svakog tijela povećamo 2 puta, baš kao i udaljenost.**

$$F_1 = G\frac{m_1 m_2}{r^2}$$
$$F_2 = G\frac{2m_1 2m_2}{(2r)^2} = G\frac{4m_1 m_2}{4r^2}$$
$$F_2 = F_1$$

4. Djeluje li gravitacijska sila Zemlje na astronauta koji se nalazi u svemirskoj stanici koja kruži oko Zemlje? Obrazložite.

Na astronauta djeluje gravitacijska sila Zemlje jer on kruži oko Zemlje zajedno sa svemirskom stanicom. Gravitacijska sila Zemlje u ulozi je centripetalne sile koja mijenja smjer brzine astronauta i svemirske stanice

5. Planet Jupiter otprilike je 300 puta masivniji od planeta Zemlje. Usprkos tome tijelo na „površini“ Jupitera ima oko tri puta veću težinu nego što ju ima na površini Zemlje. Objasnite zašto je tako mala razlika u težinama tijela jednake mase na Jupiteru i Zemlji.

$$
\begin{aligned}
M_J &= 300 M_Z \\
R_J &= 11 R_Z \\
F_{\text{na Zemlji}} &= G\frac{mM_Z}{R_Z^2} \\
F_{\text{na Jupiteru}} &= G\frac{mM_J}{R_J^2} \\
\frac{F_{\text{na Jupiteru}}}{F_{\text{na Zemlji}}} &= \frac{\cancel{G}\frac{\cancel{m}M_J}{R_J^2}}{\cancel{G}\frac{\cancel{m}M_Z}{R_Z^2}} = \frac{M_J R_Z^2}{M_Z R_J^2} \\
\frac{F_{\text{na Jupiteru}}}{F_{\text{na Zemlji}}} &= \frac{300\cancel{M_Z}\cdot\cancel{R_Z^2}}{\cancel{M_Z}\cdot 11^2\cdot\cancel{R_Z^2}} = \frac{300}{121} \\
F_{\text{na Jupiteru}} &= \frac{300}{121}F_{\text{na Zemlji}} \approx 2,5 F_{\text{na Zemlji}}
\end{aligned}
$$

Površina Jupitera je 11 puta udaljenija od njegovog centra mase nego Zemljina površina od njenog centra mase. Zbog ogromnog radijusa Jupitera, gravitacijska sila na tijela na njegovoj površini tek je oko 2,5 puta jača nego Zemljina na njenoj površini.

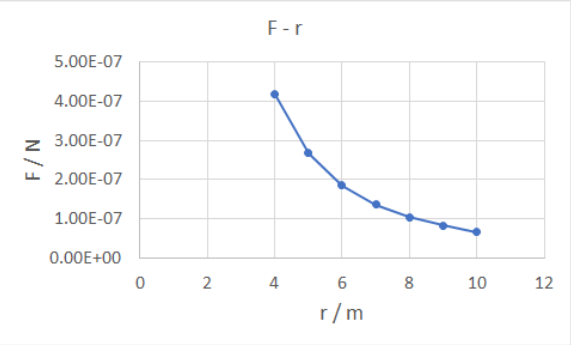
Newtonov zakon gravitacije

Ako je masa Mjeseca $m = 7{,}347 \cdot 10^{22}$ kg, kolikom silom Zemlja privlači Mjesec?

$$F_g = m \cdot g = 7{,}347 \cdot 10^{22}\, kg \cdot 9.81 \frac{N}{kg} = 7{,}21 \cdot 10^{23} N$$

IP1: Kako gravitacijska sila ovisi o udaljenosti između dvaju tijela?

r / m	F_{12} / N
10	$6,67 \cdot 10^{-8}$
9	$8,24 \cdot 10^{-8}$
8	$1,04 \cdot 10^{-7}$
7	$1,36 \cdot 10^{-7}$
6	$1,85 \cdot 10^{-7}$
5	$2,67 \cdot 10^{-7}$
4	$4,17 \cdot 10^{-7}$



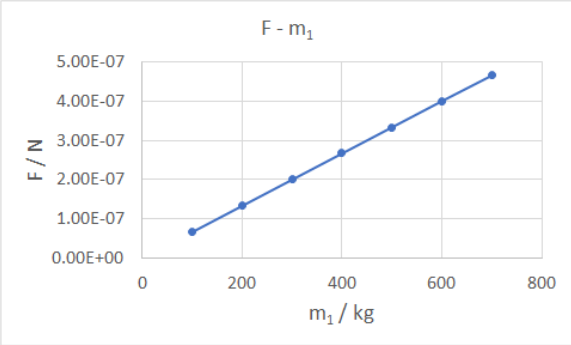
$$F \sim \frac{1}{r^2}$$

Gravitacijska sila između dva tijela pada s udaljenosti između tijela.

Gravitacijska sila je inverzno proporcionalna s kvadratom udaljenosti tijela

IP2: Kako gravitacijska sila ovisi o masi jednog od tijela?

m_1 / kg	F_{12} / N
100	$6,67 \cdot 10^{-8}$
200	$1,34 \cdot 10^{-7}$
300	$2 \cdot 10^{-7}$
400	$2,67 \cdot 10^{-7}$
500	$3,34 \cdot 10^{-7}$
600	$4 \cdot 10^{-7}$
700	$4,67 \cdot 10^{-7}$



$$F \sim m_1 \qquad F \sim m_2$$

Oblik dijagrama je pravac. Gravitacijska sila je proporcionalna s masom tijela

Newtonov zakon gravitacije:

$$F = G\frac{m_1 m_2}{r^2}$$

$G = 6,67 \cdot 10^{-11} \frac{\mathrm{Nm}^2}{\mathrm{kg}^2}$ univerzalna gravitacijska konstanta

Koristeći Newtonov zakon gravitacije može se izračunati sila Zemlje na mjesec koja odgovara tabličnoj vrijednosti:

$$F = G\frac{m_Z m_M}{r^2} = 6,67428{\cdot}10^{-11} \frac{\mathrm{Nm}^2}{\mathrm{kg}^2} \frac{5,972 \cdot 10^{24} \cdot 7,347 \cdot 10^{22}\ \mathrm{kg}^2}{384401000^2\ \mathrm{m}^2} = 1,98{\cdot}10^{20}\ \mathrm{N}$$

Slika 6.7: Plan ploče za nastavni sat o Newtonovom zakonu gravitacije.

Literatura

[1] NASA exoplanet exploration, https://exoplanets.nasa.gov/, 14. 5. 2021.

[2] Kepler mission overview, https://tinyurl.com/vadc9s78, 14. 6. 2021.

[3] Exoplanet Hunting Using Machine Learning, https://tinyurl.com/w4mw485j, 1. 7. 2021.

[4] TESS, the Transiting Exoplanet Survey Satellite, https://tess.mit.edu/, 14. 6. 2021.

[5] Planet Hunters TESS, https://tinyurl.com/krnwp2xd, 14. 6. 2021.

[6] Five ways to find a planet, https://tinyurl.com/2kzr4a5j, 14. 6. 2021.

[7] Exploring Exoplanets with Kepler, https://tinyurl.com/2nzaysbj, 14. 6. 2021.

[8] Kepler High Level Science Products (HLSP), https://tinyurl.com/pxh76s7m, 14. 6. 2021.

[9] Mitchell, T. M. Machine Learning. New York: McGraw-Hill, 1997.

[10] Géron, A. Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems. 2nd ed. Sebastopol: O'Reilly Media, 2019.

[11] Reinsel, D.; Gantz, J; Rydning, J.: The Digitization of the World, From Edge to Core, https://tinyurl.com/5yds4pfy, 7. 5. 2021.

[12] Muller, A. C.; Guido, S. Introduction to Machine Learning with Python. Sebastopol: O'Reilly Media, 2017.

[13] McCullagh, P.; Nelder, J. A. Generalized Linear Models, Second Edition. Chapman and Hall, 1989.

[14] O Pythonu, https://tinyurl.com/bpaan44a, 26. 4. 2021.

[15] O Jupyter bilježnicama, https://jupyter.org/index.html, 26. 4. 2021.

[16] O NumPy modulu, https://tinyurl.com/68hzc7fw, 26. 4. 2021.

[17] O scikit-learn modulu, https://scikit-learn.org/stable/, 26. 4. 2021.

[18] O pandas modulu, https://pandas.pydata.org/about/index.html, 27. 4. 2021.

[19] O matplotlib modulu, https://matplotlib.org/stable/index.html, 27. 4. 2021.

[20] O seaborn modulu, https://seaborn.pydata.org/, 27. 4. 2021.

[21] NASA Exoplanet Archive, https://exoplanetarchive.ipac.caltech.edu, 8. 6. 2021.

[22] Mikulski Archive, https://archive.stsci.edu/missions-and-data/k2, 8. 6. 2021.

[23] Podaci intenziteta svjetlosti zvijezda, https://tinyurl.com/my6adh46, 8. 6. 2021.

[24] Savitzky, A.; Golay, M. J. E., Smoothing and Differentiation of Data by Simplified Least Squares Procedures. // Analytical Chemistry. 1964. 36(8), str. 1627-1639

[25] Wijaya, C. Y. 5 SMOTE Techniques for Oversampling your Imbalance Data, https://tinyurl.com/dxzztnr8, 9. 6. 2021.

[26] Citizen Science, https://en.wikipedia.org/wiki/Citizen_science, 30. 6. 2021.

[27] Volunteer computing, https://tinyurl.com/ahjs6kvt, 30. 6. 2021.

[28] List of distributed computing projects, https://tinyurl.com/4hyycnay, 30. 6. 2021.

[29] Berkeley SETI Research Center, Finding artificial signals, https://tinyurl.com/hvue77bs, 30. 6. 2021.

[30] BOINC, Compute for Science, https://boinc.berkeley.edu, 30. 6. 2021.

[31] Zooniverse, People-powered research, https://www.zooniverse.org, 30. 6. 2021.

[32] Einstein@Home, https://einsteinathome.org, 30. 6. 2021.

[33] Einstein@Home, new discoveries and detections of known pulsars in the BRP4 search, https://tinyurl.com/aku6z4, 30. 6. 2021.

[34] Transiting Exoplanet Survey Satellite (TESS), https://tinyurl.com/d4u54jc4, 1.7.2021.

[35] Eisner, N.; et al., Planet Hunters TESS I: TOI 813, a subgiant hosting a transiting Saturn-sized planet on an 84-day orbit // Monthly Notices of the Royal Astronomical Society. 2020. 494(1), str. 750–763

[36] Eisner, N. et al., Planet Hunters TESS II: findings from the first two years of TESS. // Monthly Notices of the Royal Astronomical Society. 2020. 501(4), str. 4669–4690

[37] Lintott, C. From planets to policy. // Future Directions for Citizen Science and Public Policy / edited by K. Cohen and R. Doubleday. Centre for Science and Policy, University of Cambridge 2021. str. 32-38.

[38] Supernova Hunters, https://tinyurl.com/2jhptmjc, 1. 7. 2021.

[39] Landau, E. Citizen Scientists Discover Two Gaseous Planets around a Bright Sun-like Star, https://tinyurl.com/53ewsbc5, 1. 7. 2021.

[40] Fizika oko nas 1 - udžbenik fizike u prvom razredu gimnazije, https://tinyurl.com/w88dhyps, 6. 7. 2021.

[41] Negovec, H., Pavlović, D.: Fizika 1 - radna bilježnica za strukovne škole, četverogodišnji program, Zagreb: Profil Klett, 2008.

[42] Young, H. D., Freedman R. A. University Physics (13th Edition), Pearson, 2012.

[43] MacDougal, D. W., Newton's Gravity: An Introductory Guide to the Mechanics of the Universe, Springer, 2012.

[44] Interaktivna simulacija Newtonovog zakona gravitacije između dva tijela, https://phet.colorado.edu/sims/html/gravity-force-lab/latest/gravity-force-lab_en.html, 6. 7. 2021.