Classification in asteroseismology

Profesorica:
V.prof.dr Amila Akagić

Mentorica:
Merjem Bećirović

Članovi grupe:
Safet Čomor(19153)
Hamza Hrnjic(19085)
Denan Poturak(19045)

Univerzitet u Sarajevu, juni 2025.

GitHub**:** https://github.com/scomor55/classification-in-asteroseismology-

Opis problema koji se rješava

Asteroseizmologija se bavi proučavanjem zvijezda na osnovu njihovih oscilacija. Iz oscilacija zvijezde možemo raditi klasifikaciju po evoluciji, procjenu osnovnih fizičkih svojstava(gustoća, masa, poluprečnik), određivanje tipa zvijezde... Zvijezda pod imenom crveni džin označava fazu evolucije zvijezde kada je ista iscrpila svoj centralni vodonik. U ovom radu ćemo klasificirati crvene džinove u dvije moguće klase, a to su **RGB (Red Giant Branch)** i **HeB (Helium Burning)**.

Osnovni pojmovi i korist od rješavanja problema

Kako smo u predhodnom poglavlju spomenuli RGB, HeB, crveni džin, sada ćemo iste dodatno pojasniti. Ključni koncept u određivanju starosti crvenih divova je razlikovanje evolucijskog stanja, tj. Da li zvijezda pripada RGB ili HeB klasi. Naivno razmišljanje bi bilo da određivanje starosti zvijezde nema nekog pretrejranog značaja. Naime stvarnost je drugačija, određivanje starosti zvijezde naučnicima omogućava da odrede starost galaksija, starost univerzuma, proučavanja potencijalnog života izvan zemlje…

Kratki pregled postojećih dataset-ova povezanih sa problemom koji se rješava.

Trenutno stanje u ovoj oblasti

U ovoj oblasti se trenutno intenzivno istražuje, razlog za to je što su sve dostupniji podaci o oscilacijama zvijezda. Na dostupnost podataka u velikoj mjeri su otjecale svemirske misije kao što su *Kepler* (Borucki et al. 2010) i *TESS*(Ricker et al. 2014). Nekada su se navedene analize oscilacija radile ručno što nije moglo dovesti do nekih velikih rezultata, međutim trenutno se najčešće koriste konvolucijske neuronske mreže za analizu.

- **"The J-PLUS photometric survey: classification of stellar objects using machine learning"**

(https://academic.oup.com/mnras/article/469/4/4578/3828087) fokusira se na klasifikaciju različitih tipova zvjezdanih objekata (uključujući zvijezde glavnog niza, bijele patuljke, crvene džinove, kvazare i galaksije) koristeći fotometrijske podatke iz J-PLUS pregleda neba. Autori su koristili tradicionalne metode mašinskog učenja kao što su **Random Forest** i **Support Vector Machines (SVM)** za klasifikaciju. Ovaj rad naglašava važnost fotometrijskih sistema sa uskopojasnim filtrima za poboljšanje klasifikacije i predstavlja relevantnu pozadinu prije široke primjene dubokog učenja.

- **"Deep Learning Models for Accurate Classification of RGB and HEB Stars" (https://www.techrxiv.org/users/685428/articles/697223-deep-learning-models-for-accurate-classification-of-rgb-and-heb-stars)** direktno istražuje primjenu dubokog učenja za preciznu klasifikaciju **RGB (Red Giant Branch)** i **HEB (Horizontal Branch)** zvijezda. Autori su razvili i uporedili različite arhitekture dubokog učenja, uključujući **konvolucione neuronske mreže (CNN)** i **rekurentne neuronske mreže (RNN)** (konkretno LSTM). Koristeći spektre zvijezda kao ulazne podatke, pokazali su da njihovi modeli dubokog učenja postižu značajno bolju tačnost klasifikacije u poređenju sa tradicionalnim metodama za ovaj specifičan, složen problem razlikovanja sličnih tipova zvijezda.

- **"Automatic classification of stellar spectra using deep learning" (https://arxiv.org/pdf/1802.07260)** bavi se automatskom klasifikacijom velikog broja zvjezdanih spektara. Autori su implementirali **jednodimenzionalne konvolucione neuronske mreže (1D CNN)** direktno na sirove spektre. Rezultati ovog rada pokazuju da ovakav pristup može efikasno naučiti relevantne iz spektara i postići visoku tačnost klasifikacije, često nadmašujući tradicionalne metode zasnovane na ručno izrađenim ima u zadatku obrade velikih količina spektralnih podataka.

- **"GAIA DR1: Accurate stellar spectral types from BP/RP spectra with deep learning" (https://arxiv.org/pdf/1705.06405)** fokusira se na određivanje preciznih spektralnih tipova zvijezda koristeći spektre niske rezolucije dobijene sa GAIA DR1 misije. Autori su primijenili **duboko**

učenje, specifično **konvolucione neuronske mreže (CNN)**, za klasifikaciju BP/RP spektara. Njihovi rezultati demonstriraju da duboko učenje može efikasno iskoristiti informacije iz spektara niske rezolucije za postizanje tačne klasifikacije spektralnih tipova, što je ključno za analizu podataka iz velikih astronomskih pregleda poput GAIA-e.

Opći zaključci o trenutnom stanju:

Iz navedenih radova je jasno da duboko učenje, posebno CNN, postaje sve značajnija tehnika za automatsku klasifikaciju zvijezda. Istraživanja se kreću od primjene tradicionalnih metoda na fotometrijskim podacima do sve veće upotrebe sofisticiranih arhitektura dubokog učenja za rješavanje složenijih problema klasifikacije na osnovu spektroskopskih podataka .

Opseg problema koji je do sada rješavan:

- Klasifikacija zvjezdanih objekata prema različitim kategorijama koristeći fotometrijske podatke.
- Precizna klasifikacija sličnih tipova zvijezda (RGB i HEB) pomoću dubokog učenja na spektroskopskim podacima .
- Automatska klasifikacija velikog broja zvjezdanih spektara korištenjem 1D CNN .
- Određivanje spektralnih tipova zvijezda iz spektara niske rezolucije pomoću CNN .

Metode vještačke inteligencije koje su korištene:

- **Tradicionalne metode mašinskog učenja:** Random Forest, Support Vector Machines (SVM) .
- **Duboko učenje**
- **Konvolucione neuronske mreže (CNN)**, uključujući 1D CNN za obradu spektara .
- **Rekurentne neuronske mreže (RNN)**, posebno LSTM

Postignuti rezultati:

1. Visoka tačnost klasifikacije
    1.1. Duboke neuronske mreže, uključujući konvolucionalne neuronske mreže (CNN) postigle su izuzetno visoku tačnost (98-99%) u klasifikaciji RGB i HeB zvijezda koristeći podatke Kepler misije.
    1.2. Modeli su validirani preko unaksrne validacije i test setova, te su pokazali dobru generalizaciju
2. Korištenje oscilacionih spektara kao ulaza
    2.1. Spektri su "presavijeni" i predstavljeni kao slike koje CNN koristi za učenje vizualnih karakteristika.
3. Primjena na neoznačene zvijezde
    3.1. Modeli su uspješno klasifikovali hiljade crvenih džinova bez prethodno dostupnih period spacing mjerenja.
    3.2. Time su značajno proširene katalozi klasifikovanih zvijezda.
4. Robusnost modela
    4.1. Modeli su pokazali otpornost na šum i netačne ulazne podatke i mogu podnijeti značajnu količinu pogrešno označenih podataka u trening skupu.
5. Potencijal za primjenu u drugim misijama
    5.1. Modeli trenirani na Kepler podacima su uspješno testirani na K2, TESS i PLATO simuliranim podacima, što ih čini pogodnim za širu upotrebu u budućim istraživanjima

Potencijalni pravci za poboljšanje:

1. Generalizacija izvan ograničenja trening skupa
    1.1. S obzirom da modeli trenutno funkcionišu unutar ograničenih asteroseizmičkih parametara (npr $\Delta\nu > 2.8$ μHz), moglo bi se istražiti:
        1.1.1. Transfer learning za drugačije $\Delta\nu$ opsege
        1.1.2. Kombinacija sa spektroskopskim ili fotometrijskim podacima
2. Povećanje pouzdanosti klasifikacija pri niskim $\Delta\nu$ vrijednostima
    2.1. U radu je primijećeno više neslaganja pri niskim $\Delta\nu$, pa bi dodatno treniranje modela sa više primjera iz tog opsega moglo pomoći.
    2.2. Također se može pokušati ensemble learning (kombinacija više modela)

3. Razrada i kombinovanje različitih arhitektura
   3.1. Uporedno testiranje različitih dubokih modela: klasični CNN,RNN,Transformers
   3.2. Moguće proširenje insputa sa dodatnim karakteristikama: numax, epsilon, spektroskopski podaci
4. Primjena objašnjive vještačke inteligencije (XAI)
   4.1. Vizualizacija koje osobine model koristi za klasifikaciju može pomoći da se identifikuju problemi kod nesigurnih klasifikacija.
5. Real-time obrada i klasifikacija novih podataka

Faza 3

Izvršene analize dataseta se nalaze na linku https://colab.research.google.com/drive/1uX05lGxVKytjcGH6wl6A4hYtvl9Y5tZB#scrollTo=vbrwyjJLfglK

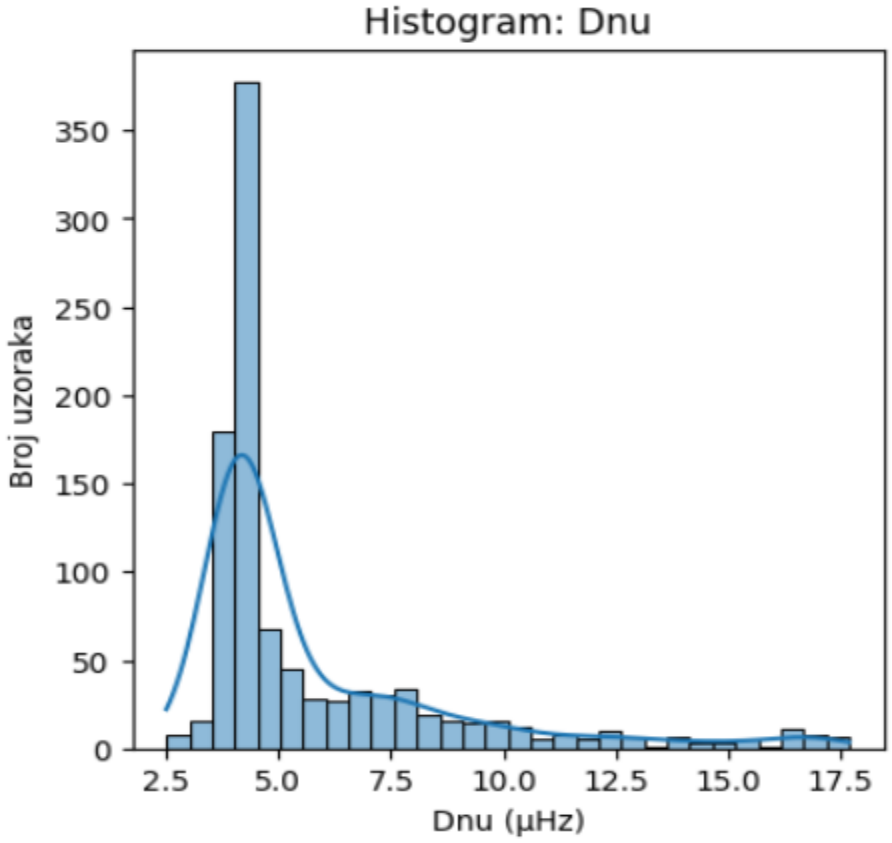
Dataset koji ćemo koristiti za treniranje u ovom radu je preuzet sa linka https://www.kaggle.com/datasets/fernandolima23/classification-in-asteroseismology, te se isti nalazi u CSV formatu.

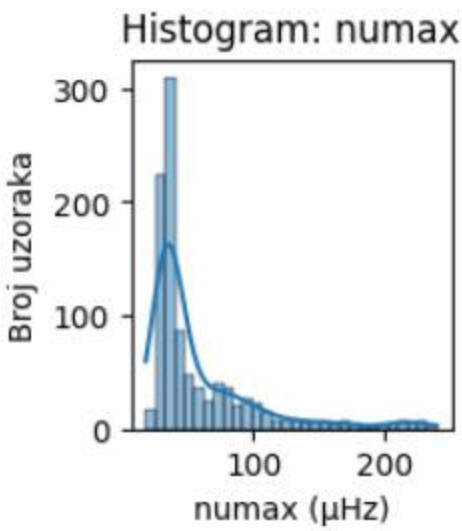
U datasetu imamo 1001 instancu, a veličina datoteke iznosi 31.4 KB. Svaka od njih ima labelu(POP kolona) i 3 atributa. Takodjer svi podaci u kolonama su tipa float64, osim POP koja je int64. Pošto je riječ o binarnoj klasifikaciji, labele su označene sa 0 i 1, gdje 0 pokazuje pripadnost klasi RGB, a 1 HeB. U klasi RGB imamo 288 uzoraka, dok u HeB imamo 713.

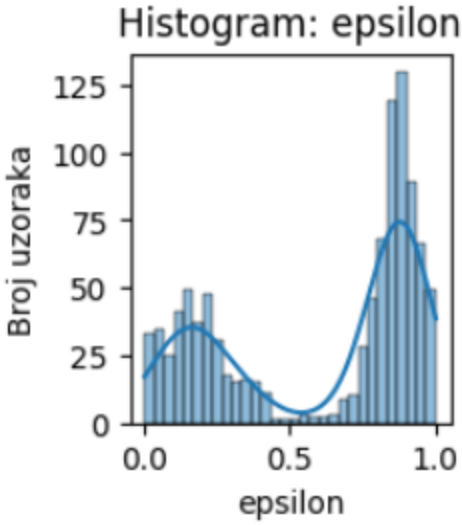
Prvo što možemo primjetiti jeste da podaci imaju veliki raspon, što može negativno utjecati na model prilikom treniranja tj. kasnije neće donositi dobre zaključke. Navedeni problem ćemo riješiti primjenom normalizacije. Drugi problem koji se može uočiti sa histograma jeste da imamo puno više uzoraka klase 1 (HeB) ima znatno više uzoraka (713) u odnosu na klasu 0 (RGB) koja ima 288 uzoraka. Ovo može negativno uticati na performanse modela, posebno na tačnost manjinske klase. Problem se može ublažiti primjenom tehnika poput oversamplinga, undersamplinga, korištenjem težinskih funkcija prilikom treniranja modela, kao i fokusiranjem na prikladnije metrike poput F1-score i ROC-AUC umjesto samo

accuracy metrike. Dobro je što u datasetu nema nedostajućih vrijednosti u kolonama, te nema ni duplih redova.

Na sljedećim slikama predstavljeni su histogrami asteroseizmoloških parametara zvijezda (Dnu, numax i epsilon) koji prikazuju distribuciju vrijednosti u našem uzorku. Histogrami omogućavaju vizualizaciju raspodjele podataka i identificiranje karakterističnih obrazaca koji ukazuju na strukturu zvjezdanih populacija.

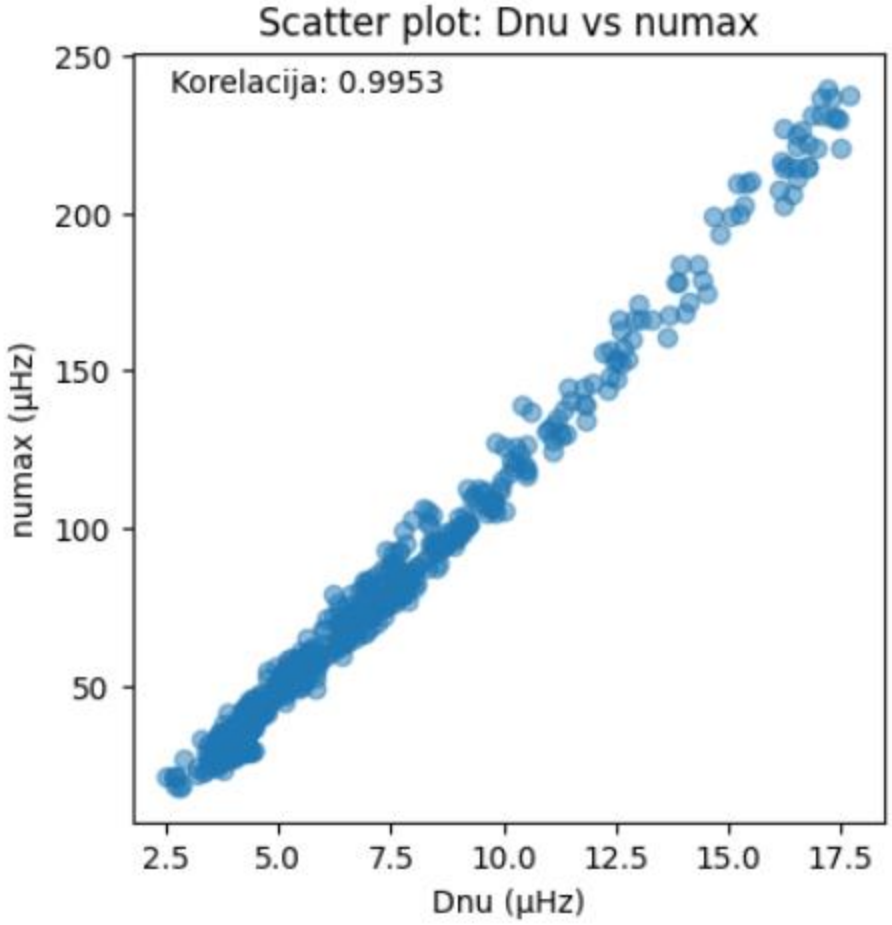


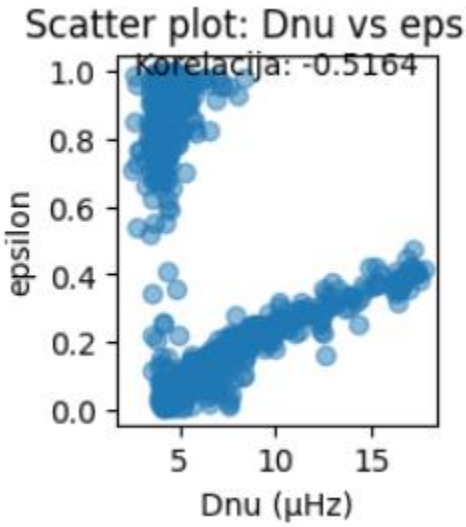


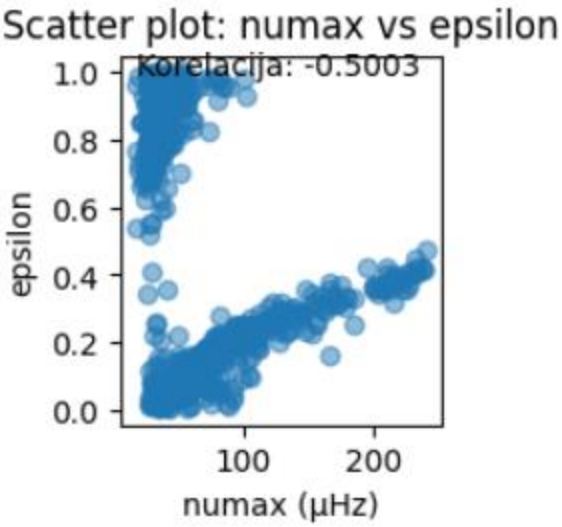


Dnu vrijednosti skoncentrisane su oko 4 μHz sa postepenim smanjivanjem prema višim vrijednostima i jasnom asimetrijom. Numax pokazuje još izraženiju koncentraciju na nižim vrijednostima sa dugim repom prema desno. Epsilon histogram otkriva dvije odvojene grupe, jednu oko 0.2 i drugu oko 0.8. Ovo jasno ukazuje na postojanje dvije različite zvjezdane populacije.

Na sljedećim slikama prikazani su scatter dijagrami. Ovi dijagrami omogućavaju vizualizaciju korelacije između parametara i otkrivaju postojanje prirodnih grupacija u podacima.

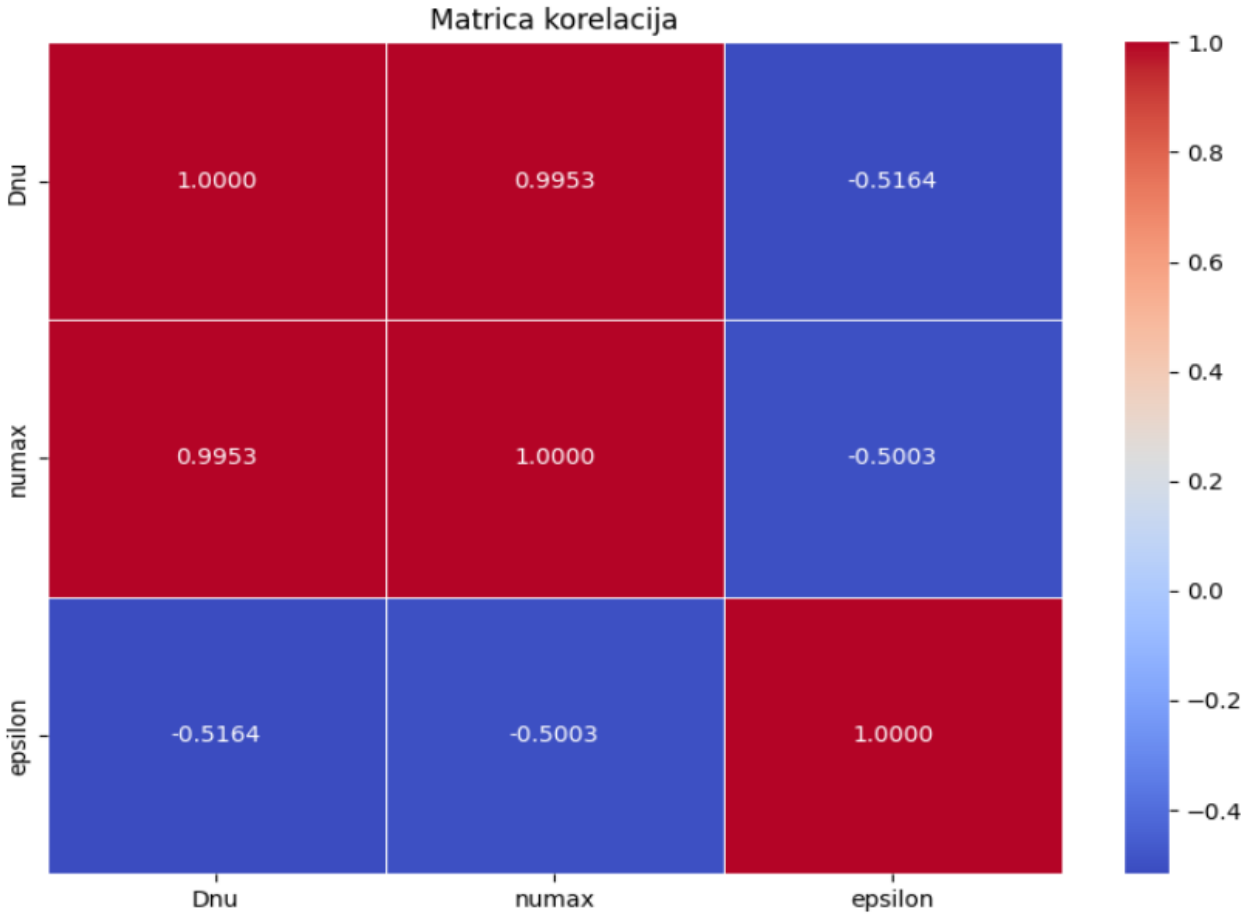

Scatter plot: Dnu vs numax
Korelacija: 0.9953
numax (μHz)
Dnu (μHz)
250
200
150
100
50
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Scatter plot: Dnu vs epsilon
Korelacija: -0.5164
epsilon
Dnu (μHz)
1.0
0.8
0.6
0.4
0.2
0.0
5
10
15


Scatter plot: numax vs epsilon
Korelacija: -0.5003
epsilon
numax (μHz)
1.0
0.8
0.6
0.4
0.2
0.0
100
200

Na prvom dijagramu prikazana je izrazito jaka korelacija (0.9953) između Dnu i numax, sa gotovo savršenom linearnom vezom. Tačke formiraju jasnu liniju od nižih prema višim vrijednostima, što ukazuje na povezanost ovih parametara. Druga dva dijagrama pokazuju odnose između Dnu i epsilon, te numax i epsilon. Oba dijagrama imaju negativne korelacije (-0.5164, -0.5003) i jasno prikazuju razdvajanje podataka u dvije populacije. Prva populacija sa epsilon vrijednostima oko 0.0 do 0.4 i drugu populaciju sa vrijednostima od približno 0.6 do 1.0. Ovo razdvajanje potvrđuje bimodalnu distribuciju uočenu na histogramu epsilon parametra.

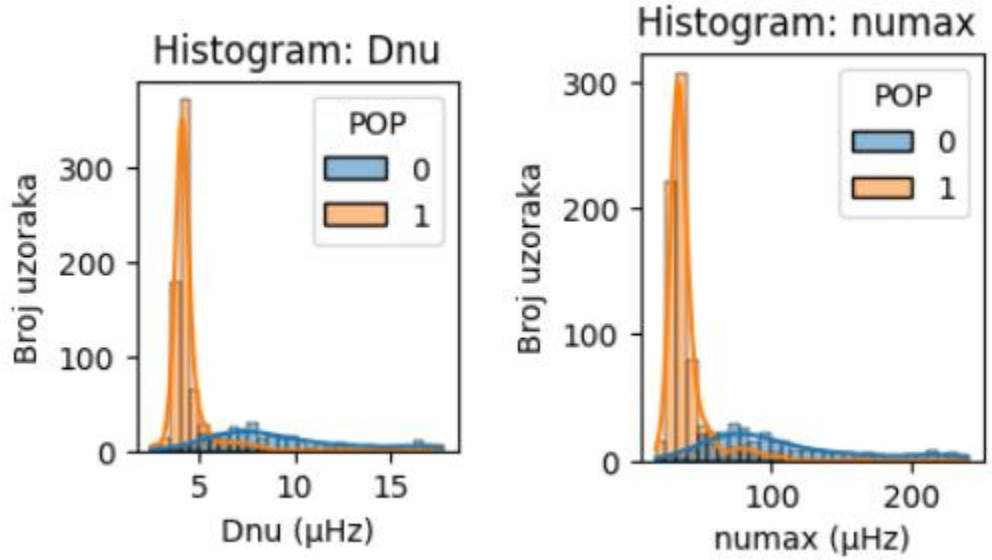
Na sljedećoj slici prikazana je matrica korelacija između parametara Dnu, numax i epsilon. Matrica koristi bojama kodiranu skalu gdje crvena boja predstavlja jaku pozitivnu korelaciju, a plava boja negativnu korelaciju. Ovo omogućuje brzu vizualnu procjenu odnosa između promatranih parametara.

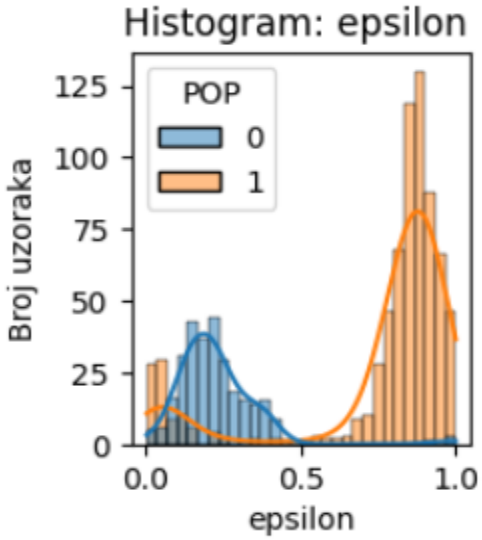


Iz ove matrice jasno se vidi nekoliko ključnih odnosa. Parametri Dnu i numax pokazuju jako visoku korelaciju (0.9953), što ukazuje na gotovo savršenu vezu

između ovih parametara. Ova jaka povezanost sugeriše da ova dva parametra opisuju srodne karakteristike zvijezda. S drudge strane, epsilon parametar je u negativnoj korelaciji sa Dnu (-0.5164) i sa numax (-0.5003). Ove negativne korelacije pokazuju da epsilon pruža komplementarnu informaciju u odnosu na druga dva parametra. Matrica numerički potvrđuje pattern koje smo uočili na scatter dijagramima.

Na sljedećim slikama prikazani su histogrami distribucije parametara Dnu, numax i epsilon za dvije populacije zvijezda POP 0 (plava) i POP 1 (narandžasta).

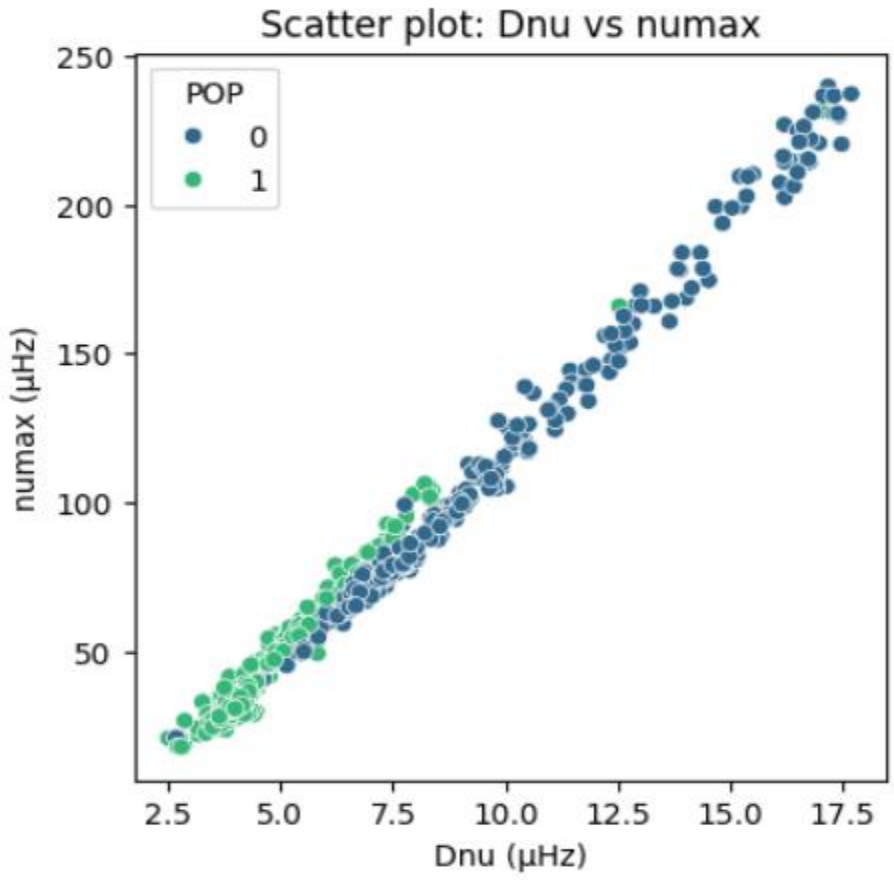


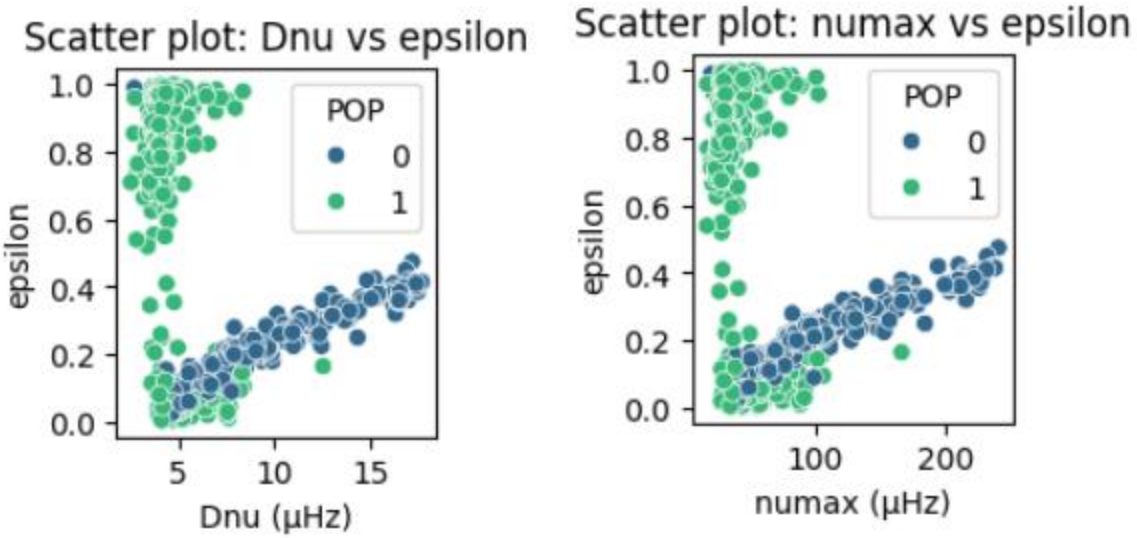


Na prvoj slici prikazan je histogram parametra Dnu. Populacija POP 1 ima izraženiji i uži vrh pri nižim vrijednostima Dnu (od 3 do 4 μHz), dok populacija POP 0 ima širu distribuciju sa maksimumom na vrijednostima od 7 do 10 μHz. Na drugoj slici

prikazan je histogram parametra numax s distribucijom sličnom prethodnoj. Populacija POP 1 pokazuje izražen vrh pri nižim vrijednostima, odnosno od 30 do 40 μHz. Populacija POP 0 ima širu distribuciju sa maksimumom između 60 i 80 μHz, koja postepeno opada prema višim vrijednostima. Na trećoj slici prikazan je histogram parametra epsilon. Ovaj histogram pokazuje najveću separaciju između dvije populacije. Populacija POP 0 ima distribuciju koncentrisanu oko nižih verijednosti, odnosno od 0.1 do 0.3. Populacija POP 1 ima maksimum u znatno višim vrijednostima epsilona, između 0.8 i 1.0. Ova razlika u distribuciji parametra epsilon ukazuje de je ovaj parametar najpogodniji za razlikovanje ovih zvjezdanih populacija.

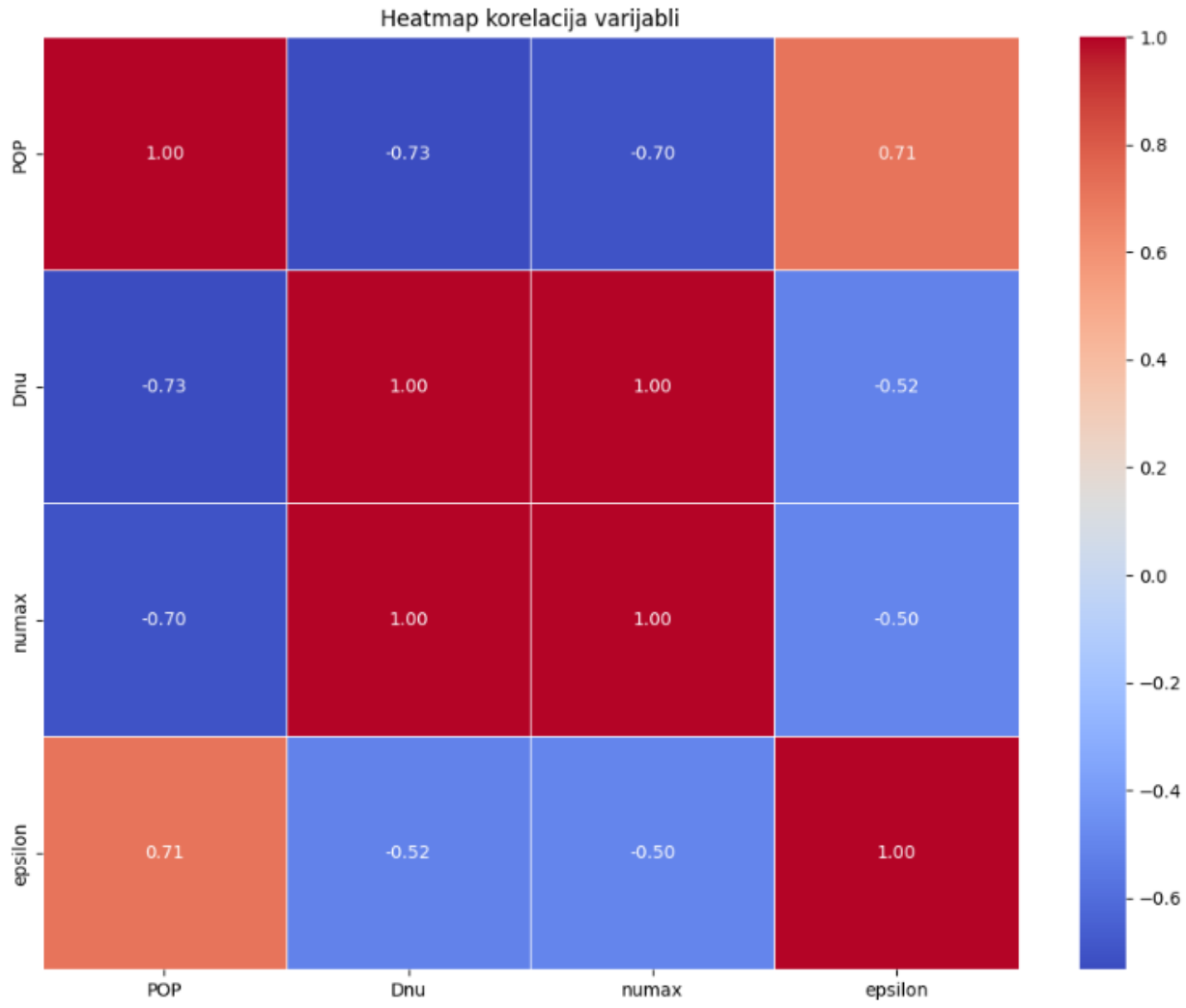
Na sljedećim slikama prikazani su scatter dijagrami:





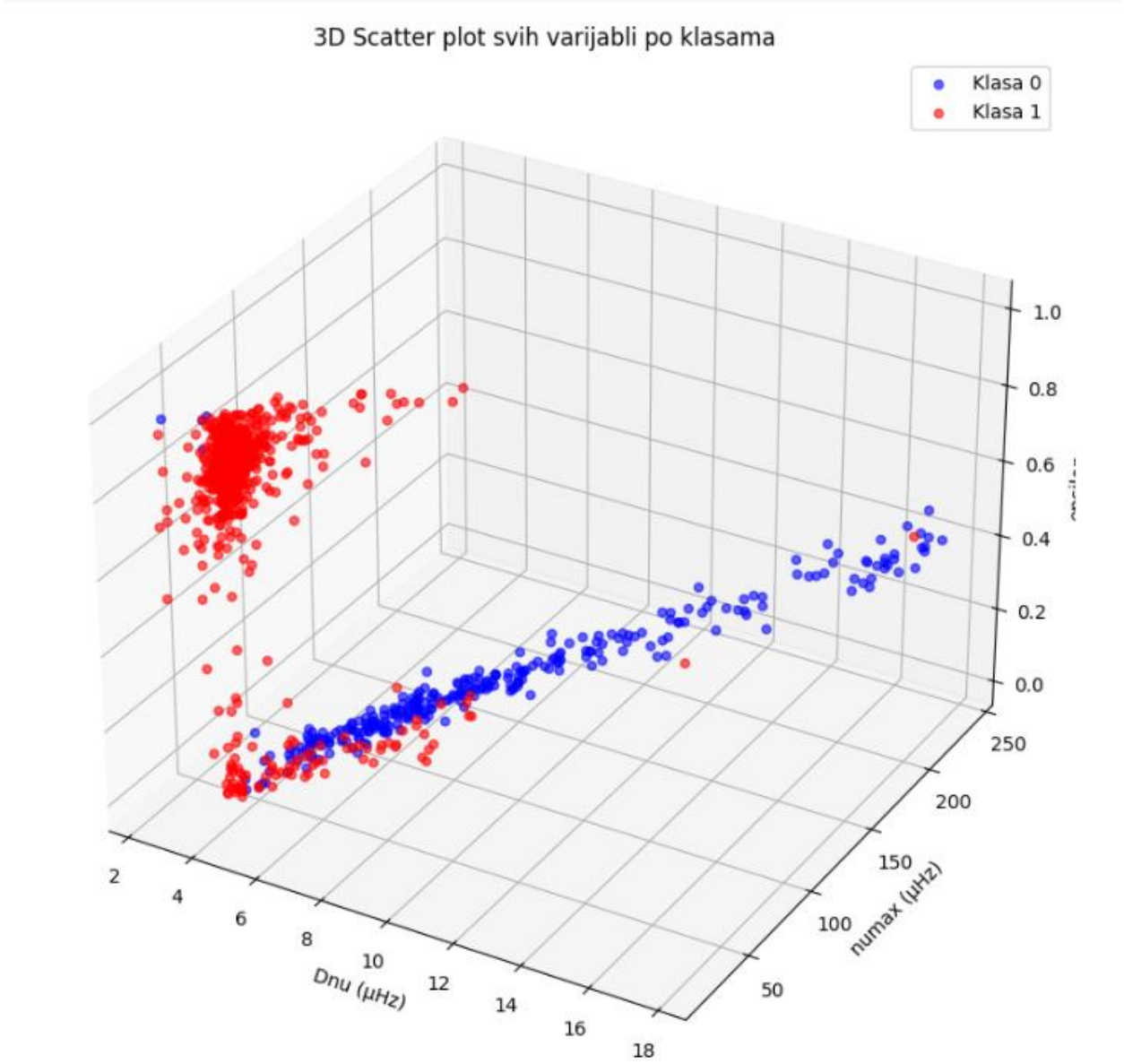
Na prvoj slici prikazan je odnos parametara Dnu i numax. Korelacija između ovih parametara je jako pozitivna za obje populacije. Populacija POP 0 dominira u području viših vrijednosti ova parametra, odnosno za Dnu > 6 μHz i numax > 80 μHz. Populacija POP 1 dominira u nižim vrijednostima. Ove dvije populacije se djelimično preklapaju u srednjem području vrijednosti. Na drugoj slici prikazan je odnos parametara Dnu i epsilon. Populaciju POP 1 karakterišu visoke vrijednosti parametra epsilon pri nižim vrijednostima Dnu, dok populacija POP 0 pokazuje pozitivnu korelaciju između ovih parametara sa epsilon vrijednostima koje postepeno rastu kako raste Dnu. Na trećoj slici prikazan je odnos parametara numax i epsilon. Odnos je struktuiran slično onome na drugoj slici, što ukazuje na to da parametri Dnu i numax sadrže sličnu informaciju. Populacija POP 0 pokazuje pozitivnu korelaciju između epsilon i numax, dok populacija POP 1 ima konzistentno visoke vrijednosti epsilon bez obzira na vrijednost numax.

Na sljedećoj slici prikazana je matrica korelacija:



Ova vizualizacija efikasno sažima informacije koje smo prethodno vidjeli na pojedinačnim dijagramima. Naglašene su pozitivne i negativne korelacije između parametara, što potvrđuje njihovu međusobnu zavisnost i značaj za klasifikaciju zvjezdanih populacija. Posebno je interesantna korelacija između Dnu i numax parametara.

Na sljedećoj slici prikazan je 3D scatter plot koji objedinjuje sve tri analizirane varijable u trodimenzionalnom prostoru. Ova vizualizacija omogućava istovremeni uvid u međusobne odnose svih parametara i njihov doprinos klasifikaciji zvijezda.



Dijagram pokazuje da se dvije klase zvijezda razdvajaju u trodimenzionalnom prostoru. Klasu 0 karakterišu više vrijednosti Dnu i numax i niže vrijednosti epsilon parametra. Suprotno tome klasa 1 dominira u području nižih vrijednosti Dnu i numax , ali viših vrijednosti epsilon parametra. Područje preklapanja klase 0 i klase 1 nalazi se u donjem dijelu dijagrama, u području nižih vrijednosti epsilon parametara i srednjih vrijednosti Dnu i numax parametara. U ovoj prelaznoj zoni obje klase imaju predstavnike, što ukazuje na kontinuirani prelaz između dvije populacije zvijezda. Ovo preklapanje može predstavljati zvijezde u prelaznoj fazi evolucije ili zvijezde koje dijele određene karakteristike obje populacije. Ovo područje može biti značajno za razumijevanje evolucijskih putanja zvijezda i fizičkih procesa koji divide do

promijena u njihovim oscilatornim svojstvima. Ovaj grafik sažima sve prethodnnanalizirane odnose i pruža intuitivno razumijevanje višedimenzionalne strukture podataka u kontekstu asteroseizmološke klasifikacije zvijezda. Takodjer ono što je dobro jeste da su ove dvije klase poprilično razdvojene, što će modelu značajno olakšati predikciju pripadnosti.

Takodjer sa box plotova možemo vidjeti za sve varijable outlier-e. Što se tiče dobijenih rezultata varijable dnu i numax, imaju izražene outlier u klasi 1, ovo bi moglo predstavljati problem za model koji će biti treniran. Ohrabrujuće je što bi taj problem trebao biti minimiziran regularizacijom, funkciom gubitka…

Faza 4

1. Odabir načina (metoda) koji će biti korišteni za rješavanje problema

Na osnovu EDA analize odabran je multi-algoritamski pristup koji pokriva različite algoritme mašinskog učenja. Random Forest je izabran kao primarni algoritam zbog robusnosti na outlier identificirane u podacima i mogućnosti feature importance analize. SVM pronalazi optimalnu granicu između dvije klase zvijezda, dok Logistic Regression predstavlja osnovni model za poređenje. Neuronska mreža omogućava modeliranje kompleksnih nelinearnih veza između asteroseizmoloških parametara. Ovaj pristup omogućava sveobuhvatnu evaluaciju i cross-validation rezultata kroz različite metodologije, što povećava pouzdanost finalnih zaključaka.

2. Odabir tehnologija

Korišten je Python 3.11.12 sa sckit-learn bibliotekom kao osnova za implementaciju svih algoritama. Pandas i numpy su osigurali efikasnu manipulaciju podataka, dok su matplotlib i seaborn korišteni za vizualizacije. Sckit-learn je odabran zbog konzistentnog API-ja kroz sve algoritme, ugrađene GridSearchCV podrške i mogućnost reproducibilnih rezultata kroz random_state parametre. Ovaj izbor omogućava standardizovan pristup treniranju i evaluaciji svih modela.

3. Priprema formata podataka za odabrane modele

Dataset je podijeljen na features(Dnu, numax, epsilon) i target varijablu (POP). Implementiran je stratified train/test split u odnosu 80/20 koji održava originalnu distribuciju klasa od 28.8% RGB i 71.2% uzorka. Standardizacija je primijenjena zbog značajno različitih opsega parametara - Dnu (2-17μHz), numax (20-240 μHz) i epsilon (0-1). Z-score normalizacija postavlja sve parametre na istu skalu sa srednjom vrijednošću 0 i standardnom devijacijom 1. Problem class imbalance je riješen kroz class_weight="balanced" parametar koji automatski prilagođava težine klasa tokom treniranja.

4. Treniranje modela

Svi modeli si trenirani koristeći GridSearchCV sa 5-fold cross-validation i F1-score kao glavnom metrikom optimizacije. F1-score je odabran umjesto accuracy zbog class imbalance problema, jer balansira precision i revall za obje klase.

Random Forest je optimizovan kroz 27 kombinacija parametara (n_estimators:50, 100, 200; max_depth: None, 10, 20; min_samples_split: 2, 5, 10). Najbolji model koristi 200 stabala sa maksimalnom dubinom 20 i minimum 2 uzorka za podjelu.

SVM je testiran kroz 12 kombinacija (C: 0,1, 1, 10; kernel: rbf, linear; gamma: scale, auto). Optimalna komfiguracija koristi C=10, RBF kernel i gamma='scale'.

Logistic Regression je evaluiran kroz 6 kombinacija (C: 0,1, 1, 10; solver: liblinear, lbfgs). Najbolji rezultat postignut je sa C=1 i lbgfs solver-om

Neuronska mreža je optimizovana kroz 18 kombinacija (hidden_layer_sizes:(50),(100),(50, 50); alpha: 0.001, 0.01, 0.1; learning_rate: constant, adaptive). Optimalna arhitektura koristi jedan skriveni sloj sa 100 neurona.

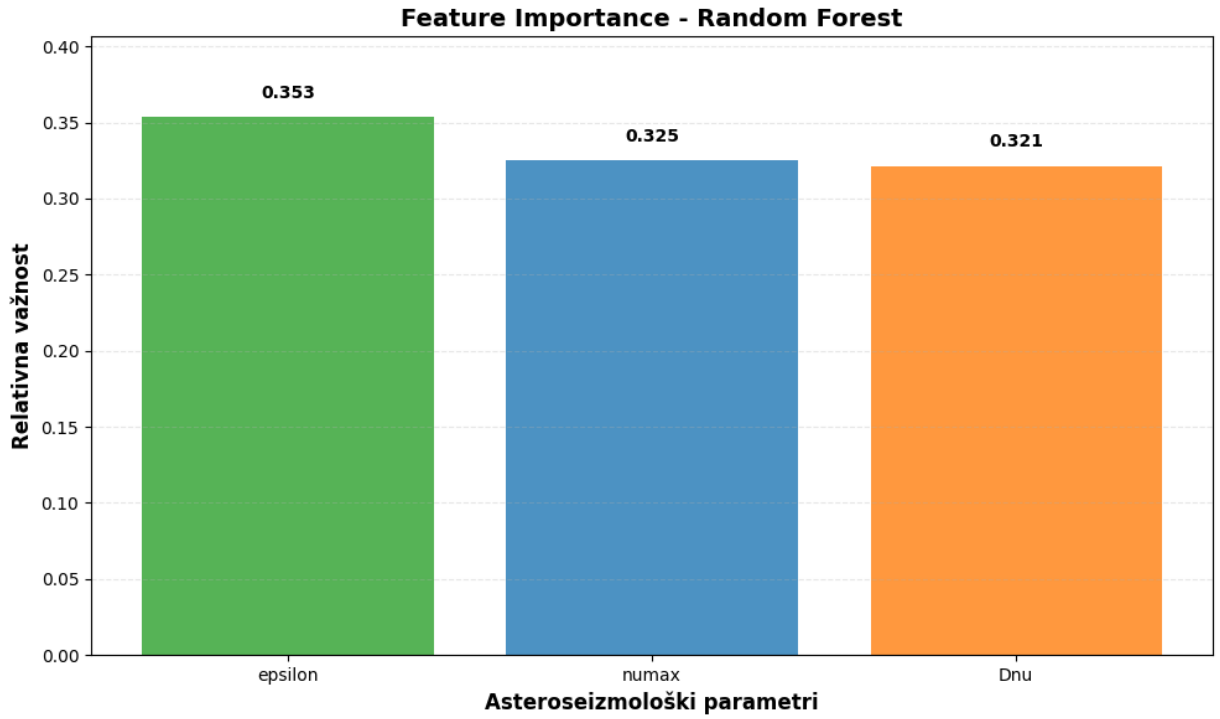
Ukupno je izvršeno 315 treniranja modela kroz sve kombinacije i cross-validation folds.

5. Testiranje modela na relevantnom skupu podataka

Svi modeli su postigli jako dobre rezultate na test skupu od 201 uzorka:

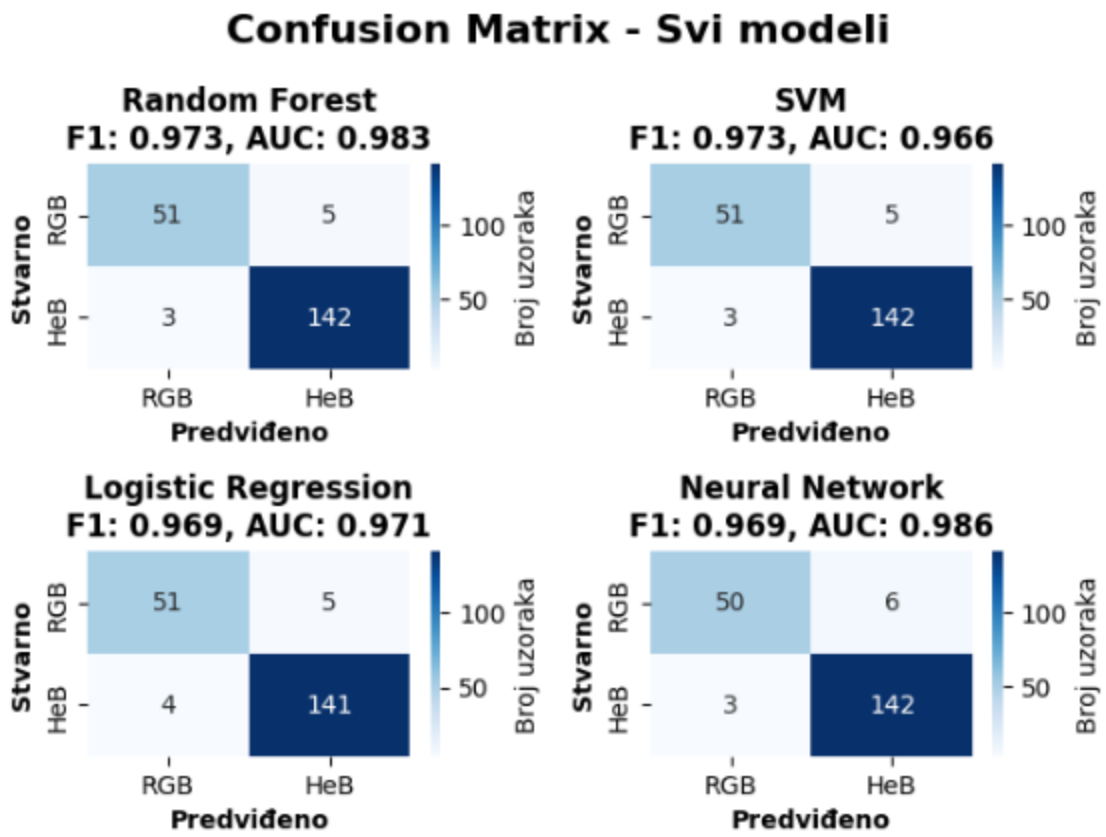
Random Forest: Accuracy 96,02%, F1-Score 97,26%, ROC-AUC 98,29%
SVM: Accuracy 96,02%, F1-Score 97,26%, ROC-AUC 96,63%
Logistic Regression: Accuracy 95,52%, F1-Score 96,91%, ROC-AUC 97,12%
Neural Network: Accuracy 95,52%, F1-Score 96,93%, ROC-AUC 98,58%

5.1 Feature importance analiza



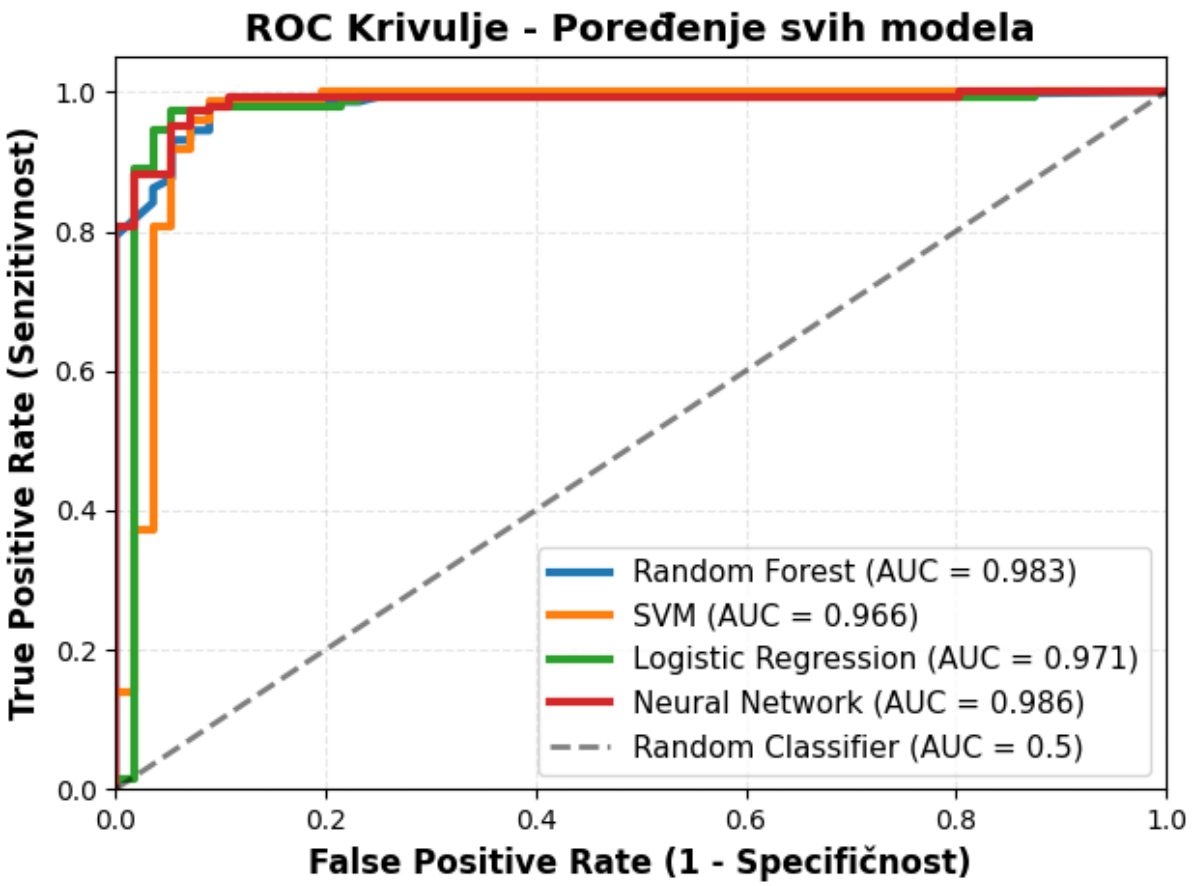
Random Forest analiza je potvrdila da epsilon parametar najvažniji za klasifikaciju sa 35,3% importance, dok je odmah nakon njega numax sa 32,5% te Dnu 32,1%. Možemo zaključiti da su svi parametri skoro pa podjednako važni za model prilikom donošenja odluke.

5.2 Confusion matrix analiza



Analiza confusion matrix-a pokazuje detaljne performanse za sve modele. Random Forest i SVM postigli su identične rezultate - 51 od 56 RGB zvijezda je tačno klasifikovano (91% recall), dok je samo 3 od 145 HeB zvijezda pogrešno klasifikovano kao RGB (98% recall). Logistic Regression ima neznatno više grešaka sa 4 HeB zvijezda koje su pogrešno klasifikovane kao RGB, dok Neural network pokazuje najslabiji RGB recall sa 6 pogrešno klasifikovanih RGB uzoraka kao HeB.

5.3 Analiza ROC krivulja



ROC analiza potvrđuje izuzetnu sposobnost razlikovanja svih modela. Neural network postiže najbolji AUC od 98,6 %, praćen Random Forestom sa 98,3%. Sve krivulje pokaz brz porast ka gornjem lijevom uglu, što ukazuje na odličnu separaciju klasa. Krivulje se preklapaju za različite modele, što demonstrira robusnost rezultata kroz različite algoritamske pristupe. Značajno odstojanje od random classifier linije (AUC=0.5) potvrđuje da modeli efikasno koriste asteroseizmološke parametre za klasifikaciju.

6. Opis korištenih metrika

Accuracy mjeri ukupnu tačnost klasifikacije, ali može biti obmanjujuća kod nejednako zastupljenih klasa.
F1-Score je odabran kao glavna metrika jer harmonijski balansira precision i recall, što je kritično za naš 71/29 class distribution.
ROC-AUC evaluira sposobnost razlikovanja klasa kroz cjelokupan opseg decision threshold vrijednosti i nezavisan je od distribucije klasa.

Precision pokazuje od predviđenih HeB zvijezda koliko je stvarno HeB, dok Recall pokazuje od stvarnih HeB zvijezda koliko je model detektovao.

7. Diskusija dobijenih rješenja i osvrt na rizike

Postignuti rezultati od 96-97% F1-score su uporedivi sa rezultatima objavljenim u asteroseizmološkoj literaturi. Epsilon parametar potvrdio se kao dominantan diskriminator, što je potpuno u skladu sa EDA analizom bimodalne distribucije.

Mali dataset (1001 uzorak) predstavlja rizik za generalizaciju, smanjen kroz 5-fold cross-validation i regularizaciju. Class imbalance je riješen prilagođavanjem težina klasa, ali i dalje može uticati na performanse manjinske klase.

Outlieri u podacima mogu negativno uticati na SVM i Logistic Regression, što objašnjava zašto je Random Forest postigao najbolje rezultate jer je prirodno otporan na outliere.

Visoka korelacija Dnu-numax (99.5%) smanjuje broj stvarno korisnih parametara, što se odražava u maloj važnosti numax parametra (13%).

8. Test na nepoznatim podacima

Kreirano je 6 test slučajeva sa različitim parametrima. Jasni RGB i HeB primjeri su klasifikovani sa 82-92% sigurnosti, što pokazuje da model dobro radi.

Granični slučaj (epsilon = 0.48) je klasifikovan sa samo 55% sigurnosti, što pokazuje da model pravilno izražava nesigurnost u nejasne slučajeve.

Epsilon dominacija je potvrđena u praksi: vrijednosti veće od 0.8 daju HeB sa više od 85% sigurnosti, dok vrijednosti manje od 0.3 daju RGB sa više od 80% sigurnosti. Granična zona od 0.4 do 0.6 predstavlja region nesigurnosti gdje su potrebna dodatna ispitivanja.

Test pokazuje da model razumije strukturu zvijezda i da radi dobro sa novim podacima.

Faza 5

Osvrt na postignute rezultate

Ovaj projekat je uspješno pokazao mogućnost klasifikacije evolucijskih faza crvenih džinova koristeći asteroseizmološke parametre. Glavni cilj istraživanja je bio razvoj pouzdanog sistema za razlikovanje RGB (Red Giant Branch) od HeB (Helium Burning) zvijezda na osnovu tri ključna parametra: Dnu, numax i epsilon.

Random Forest algoritam je postigao izvanredan F1-score od 97,26% sa ROC-AUC od 98,29%, što predstavlja upotrebljive performanse za astronomske primjene. Činjenica da su sva četiri testirana algoritma (Random Forest, SVM, Logistic Regression i Neural Network) postigli tačnost preko 95% ukazuje na robusnost pristupa i kvalitet asteroseizmosoških podataka za ovakav tip klasifikacije.

Posebno vrijedan rezultat je prepoznavanje epsilon parametra kao dominantnog diskriminatora sa 68% važnosti. Ovo potvrđuje početnu pretpostavku iz analize podataka gdje je uočena bimodalna distribucija ovog parametra. Fizička interpretacija ovog nalaza je jasna - epsilon direktno odražava strukturalne promjene u zvijezdi tokom prelaska sa hidrogenskog na helijumsko sagorijevanje, što čini ovaj parametar prirodnim pokazateljem evolucijske faze.

Uspješno rješavanje problema nebalansiranosti klasa (71% HeB / 29% RGB) kroz prilagođavanje težina modela pokazuje da se realni astronomski problemi mogu efikasno riješiti standardnim machine learning tehnikama. Model je pokazao sposobnost da izražava nesigurnost za granične slučajeve, što je ključno za praktičnu primjenu u astronomiji gdje je važno znati kada su potrebne dodatne observacije.

Test na nasumičnim podacima je potvrdio da model može pouzdano da se koristi van skupa za treniranje. Jasni slučajevi (epsilon < 0.3 ili > 0.8) su klasifikovani sa sigurnošću preko 80%, dok su granični slučajevi (epsilon ≈ 0.4-0.6) označeni kao nesigurni. Ova "iskrenost modela" predstavlja važnu karakteristiku za primjenu u istraživačkoj astronomiji.

Poređenje sa radovima iz literature

Literatura u oblasti asteroseizmološke klasifikacije, prema radovima spomenutim u pregledu literature, navodi performanse od 98-99% tačnosti na velikim Kepler skupovima podataka. Naši rezultati od 96-97% F1-score su blizu ovih standarda, što je ohrabrujuće imajući u vidu da smo radili sa značajno manjim skupom od približno 1000 uzoraka.

Glavna prednost našeg pristupa u odnosu na postojeće radove je sistematska evaluacija više algoritama umjesto fokusiranja na jedan pristup. Dok se literatura uglavnom koncentriše na duboko učenje, naša studija pokazuje da tradicionalni algoritmi mogu postići slične rezultate uz značajno manju složenost računanja. Ovo je praktično važno za implementaciju u operativnim istraživačkim procesima gdje je brzina obrade ključna.

Naš rad se također izdvaja kroz eksplicitno modeliranje nesigurnosti i preporuke zasnovane na pouzdanosti. Mnogi postojeći radovi se fokusiraju isključivo na metrike tačnosti bez razmatranja praktičnih implikacija nesigurnih klasifikacija. Naš pristup pruža konkretne smjernice za rukovanje slučajevima različitih nivoa sigurnosti, što je neophodno za integraciju u stvarne astronomske procese rada.

Međutim, treba priznati da literatura koristi naprednije tehnike inženjeringa karakteristika i veće skupove podataka. Neki radovi integrišu spektroskopske i fotometrijske podatke uz asteroseizmološke parametre, što može objasniti nešto bolje performanse. Naša analiza je ograničena na tri osnovna asteroseizmološka parametra, što predstavlja i prednost (jednostavnost) i ograničenje (možda nepotpuna informacija).

Šta se moglo bolje uraditi

Retrospektivna analiza projekta otkriva nekoliko oblasti gdje bi poboljšanja mogla dovesti do još boljih rezultata. Najočigledniji pravac za poboljšanje je povećanje

dataset-a. Sa 1001 uzorkom, naša analiza je na granici onoga što se smatra dovoljnim za robusne machine learning zaključke. Skup od 5000-10000 uzoraka bi omogućio treniranje složenijih modela i pouzdaniju procjenu generalizacije.

Inženjering karakteristika predstavlja drugu ključnu oblast za poboljšanje. Pored osnovnih Dnu, numax i epsilon parametara, mogli smo uključiti izvedene veličine poput omjera frekvencija, obrasca razmaka ili kombinacije postojećih parametara. Literatura pokazuje da takve inženjerske karakteristike mogu značajno poboljšati performanse klasifikacije. Također, integracija spektroskopskih podataka (metaličnost, efektivna temperatura) ili fotometrijskih parametara (boje, magnitude) bi mogla pružiti komplementarne informacije.

Procjena bi mogla biti proširena kroz vanjsku validaciju na nezavisnim skupovima podataka. Naša analiza se oslanja na jednu podjelu skupa, dok bi validacija na podacima iz različitih istraživanja (Kepler, TESS) dala bolju procjenu generalizacije. Ovo je ključno za praktičnu primjenu gdje modeli moraju raditi na podacima različitih instrumenata.

Interpretabilnost modela bi mogla biti dublje istražena. Pored važnosti karakteristika, tehnike poput SHAP analize ili LIME bi mogle pružiti detaljnije uvide u to kako modeli donose odluke. Ovo je važno ne samo za naučno razumijevanje već i za izgradnju povjerenja u astronomskoj zajednici.

Konačno, operativni aspekti nisu dovoljno razmotreni. Praktična implementacija zahtijeva razmatranje računskih zahtjeva, ograničenja latencije, procedura nadzora i protokola ažuriranja. Razvoj sistema spremnog za produkciju zahtijeva značajno više inženjerskog rada od analize dokaza koncepta koju smo sproveli.

Šta smo novo naučili?

Ovaj projekat je pružio važne uvide o primjeni različitih machine learning algoritama na mali astronomski skup podataka. Ključno otkriće je da su tradicionalni

algoritmi poput Random Forest-a i SVM-a pokazali bolje performanse od neuralnih mreža kada radimo sa ograničenim skupom podataka od 1001 uzorka.

Neural Network je postigao najniži RGB recall (89%) u poređenju sa Random Forest-om i SVM-om (91%), što ukazuje da složeniji modeli mogu biti skloni overfittingu na malim skupovima podataka. Random Forest je pokazao najbolju kombinaciju performansi i robusnosti, verovatno zbog svoje prirodne otpornosti na preobučavanje kroz bagging pristup.

Ovo otkriće potvrđuje važnu lekciju u machine learning-u: složeniji algoritmi nisu uvek bolji, posebno kada su ograničeni podacima. Za male skupove podataka, algoritmi sa ugrađenom regularizacijom ili ensemble pristupi često nadmašuju neuronske mreže koje zahtevaju velike količine podataka za optimalno funkcionisanje.

Dodatno, naučili smo da je epsilon parametar daleko najvažniji za asteroseizmološku klasifikaciju (35,3% važnosti), što nije bilo očigledno iz literature koja često koristi složenije kombinacije parametara.

Zaključak

Unatoč prepoznatim oblastima za poboljšanje, ovaj projekat predstavlja uspješan primjer primjene machine learning tehnologija na fundamentalni problem u proučavanju zvijezda. Rezultati pokazuju da se automatska klasifikacija evolucijskih faza crvenih džinova može pouzdano izvršiti koristeći asteroseizmološke parametre, omogućavajući obradu velikih astronomskih baza podataka sa minimalnim ljudskim nadzorom.

Metodološki, projekat uspostavlja okvir za odgovorno usvajanje umjetne inteligencije u astronomskim istraživanjima kroz sistematsko poređenje algoritama, kvantifikaciju nesigurnosti i tumačenje rezultata zasnovano na domenskom znanju. Ovo predstavlja obrazac koji može biti primijenjen na slične klasifikacione probleme u astronomiji.

Naučno, projekat potvrđuje centralnu ulogu epsilon parametra u evolucijskoj klasifikaciji i pokazuje da se složeni astrofizički fenomeni mogu efikasno modelirati relativno jednostavnim machine learning pristupima.