UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO
ODDELEK ZA FIZIKO

PRAKTIKUM STROJNEGA UČENJA V FIZIKI

2. naloga: Klasifikacija zvezdnih spektrov

Žiga Šinigoj, 28222025

Ljubljana, november 2023

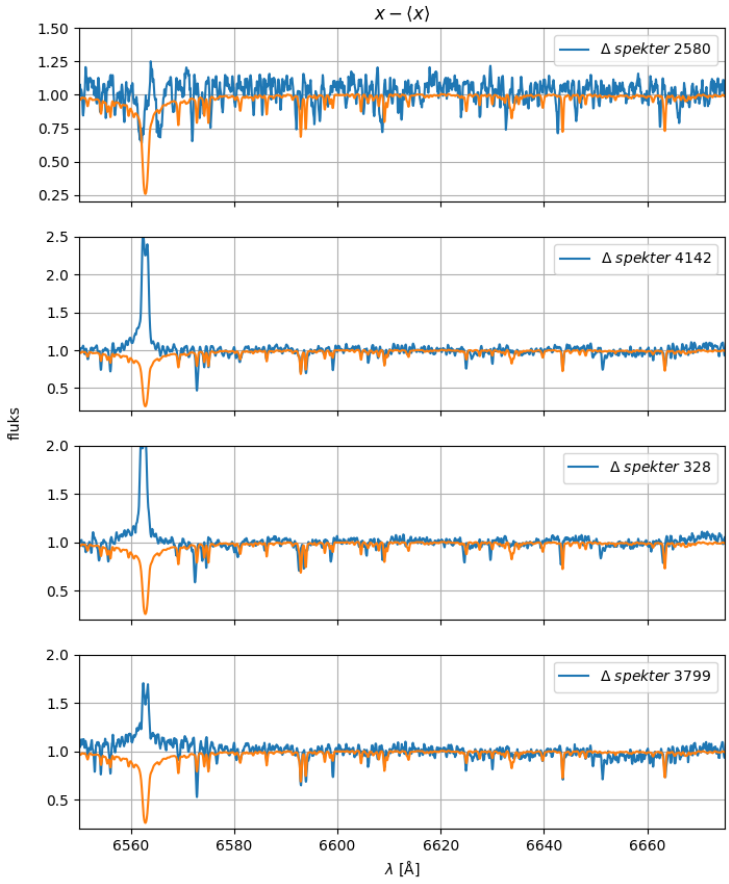
1 Uvod

Pri opazovanju zvezd lahko dobimo veliko informacij o lastnostih zvezde iz njenega spektra. Na obliko spektra vpliva temperatura, zastopanost različnih elementov - kovinskost $[M/H]$, gravitacijski pospešek, podan kot *logg*, radialna hitrost, rotacijska hitrost, turbulenca, magnetna polja, ... Ker je modeliranje spektra kompliciranо, je dobro uporabiti metode strojnega učenja. Ideja je uporabiti strojno učenje, da ločimo zvezdne spektre med samo in s tem najdemo različne tipe zvezd. Kot povsod v strojnem učenju je pomembno da imamo velik nabor podatkov za naš model, pri klasifikaciji zvezd je to koristno zaradi podobnosti nekaterih spektrov, ki ne spadajo v isti razred. Cilj naloge je najti (fizikalno) različne razrede zvzednih spektrov. V nalogi uporabimo metodo PCA, t-SNE in DBSCAN iz knjižnice scikit-learn.

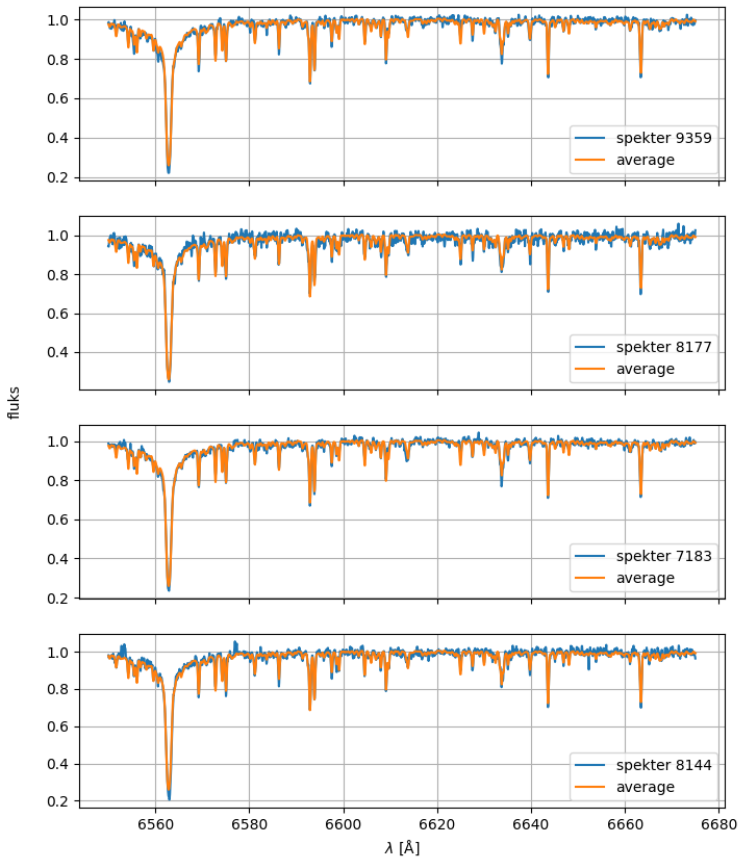
Mogoče bi bilo koristno za naprej, da se doda v virtualno okolje na marvin serverju knjižnico corner, za lažje risanje kotnih grafov (trenutno ni nameščena).

2 Podatki in odstopajoči spektri

Dobljeni set podatkov je vseboval učni set parametrov in fizikalnih količin nekaterih spektrov, s katerimi smo si lahko pomagali pri klasifikaciji. Najprej je bilo potrebno najti spektre, ki najbolj odstopajo od povprečja po nekem kriteriju oz. metriki. Najprej sem vzel za kriterij evklidsko metriko, a se ni izkazala za najboljšo metriko. Ker so spektri zašumljeni je mogoče smiselno vzeti kar razliko med spektrom in povprečjem in na ta način dobiti nenavadne spektre, označil jih bom z O1 (slika 1). Za drugo skupino nenavadni spektrov sem vzel tiste, ki so najbolj odstopali od povprečja v PCA prostoru po evklidski metriki v 100 dimenzijah, te bom označil z O2 (slika 2). Spektri O1 večinoma izstopajo po močni emisiji, medtem ko spektri O2 niso preveč izstopajoči, mogoče vsebujejo samo več šuma od povprečnega spektra. V nadaljnih izračunih sem vzel malo več O1 spektrov (10).



Slika 1: Najbolj izstopajoči spektri O1.

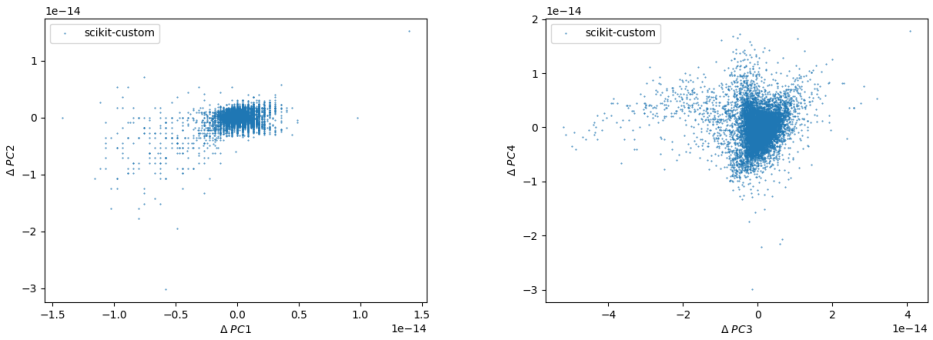


(a)

Slika 2: Najbolj izstopajoči spektri (O2) glede na 100 prvih PCA komponent.

3 Metoda PCA in vizualizacija

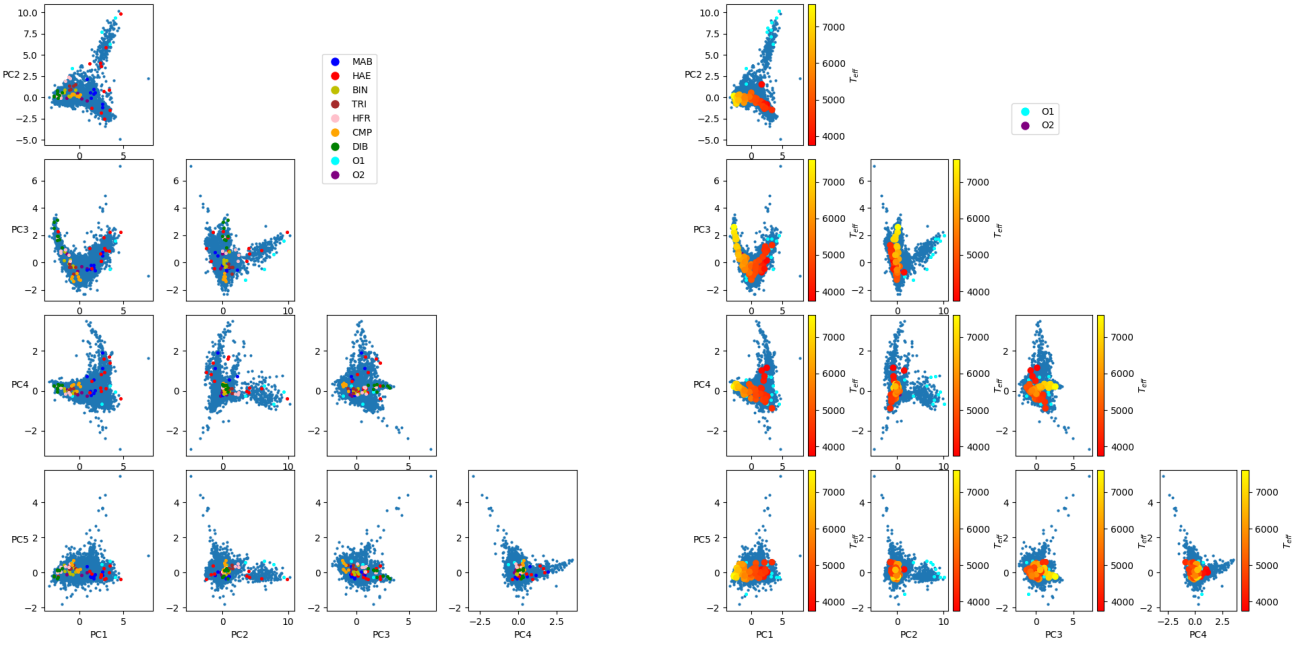
Metoda PCA je linearna metoda, katere namen je projekcija n-dimenzionalnega prostora $n >> 1$ prostor manjših dimenzij m. Večinoma nas zanima projekcija vzdolž smeri z največjimi lastnimi vektorji, saj ti predstavljajo smeri največje variance. Dobljene smeri niso nujno fizikalne količine, ampak so v splošnem njihova linearna kombinacija. Ideja je prepoznati kakšne fizikalne količine oz. korelacije med njimi s pomočjo projekcij. Implementacija PCA metode je priložena na koncu poročila. Odstopanje moje implementacije od metode iz scikit-learn je minimalno (slika 3,4). Pri vizualizaciji glavnih komponent sem naredil kotni



(a) (b)

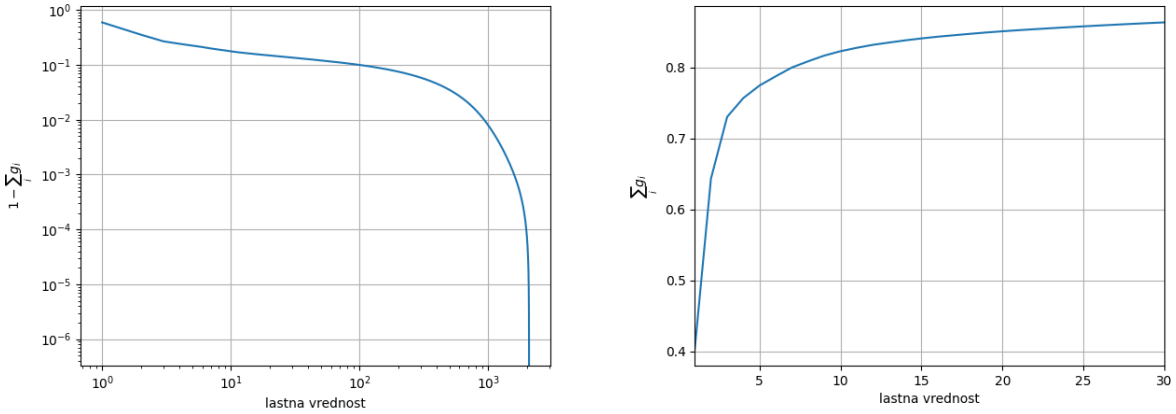
Slika 3: a) Odstopanje moje implementacije PCA in scikit-learn-ove implementacije v prvih dveh komponentah. b) Odstopanje moje implementacije PCA in scikit-learn-ove implementacije v tretji in četrti komponenti.

diagram prvih petih glavnih PCA komponent. Na dane rezultate sem narisal enkrat učne tipe in drugič učne spektre z danim fizikalnim parametrom (slike 4a, 4b, 6). Opazim lahko, da so spektri O2 precej odvisni od PCA2 komponente in, da učni parametri ne opisujejo temperatur teh zvezd. Opazim tudi, da je temperatura najbolj odvisna od parametra PC1, kar namiguje na to, da bi lahko bil v veliki večini parameter PC1 efektivna temperatura. Drugi večinski del korelacije s komponente PC1 predstavlja količina *logg*. Ostale fizikalne količine ne morem tako enostavno povezati s projekcijami komponent, saj so njihova linearna kombinacija. Opazim tudi, da si spektri O1 v bližini tipa zvezd HAE. Glede spektrov O2 pa ni kakšnih posebnosti, kar namiguje na to, da so mogoče samo bolj zašumljeni že znani spektri. Ker spektri O1 precej odstopajo sem jih primerjal s spektri HAE in povprečnim spektrom iz območja $PCA2 \geq 3.5$ (slika 7). Opazim lahko, da najverjetneje ne gre za nov tip zvezd, ampak so O1 samo spektri z veliko večjo emisijsko črto kot HAE. Na tem območju se večinoma nahajajo spektri HAE.



(a) (b)

Slika 4: a) PCA kotni diagram z dodanimi učnimi tipi in spektri O1, O2. b) PCA kotni diagram z dodanimi učnimi seti temperatur in spektri O1, O2.

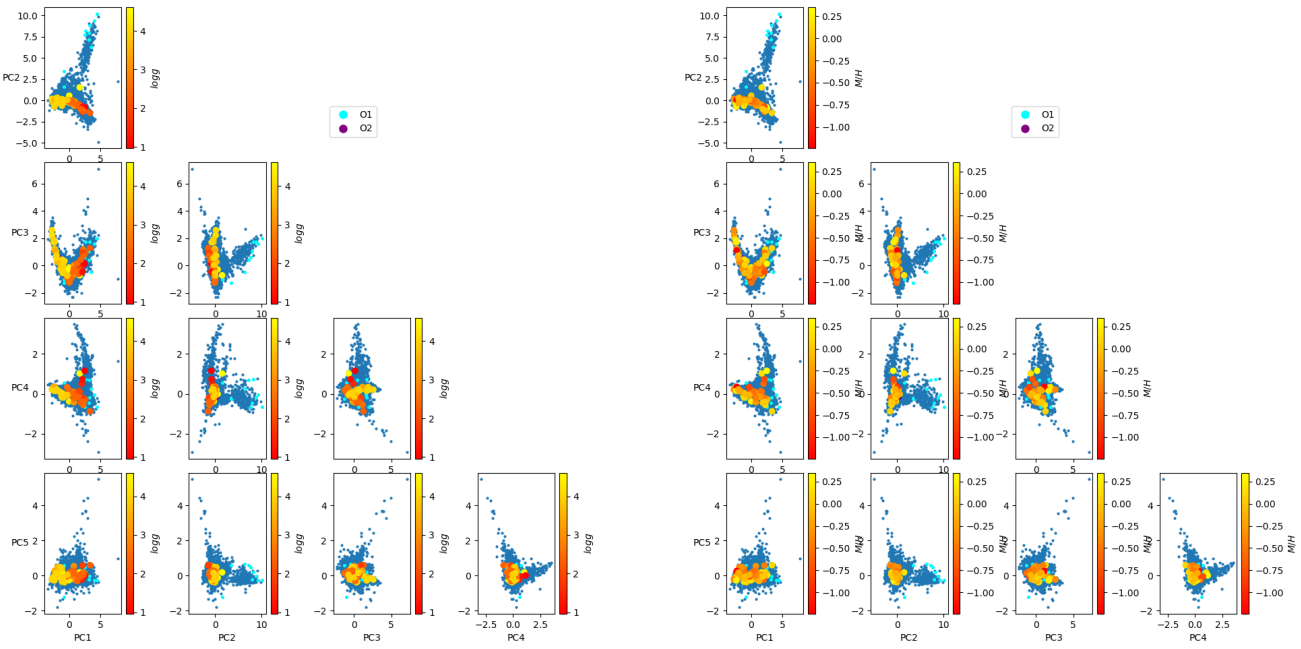


(a) (b)

Slika 5: a) Odstopanje od celotne energije pri dodajanju lastnih vrednosti h komulativni energiji. b) Komulativna energija v odvisnosti od največjih lastnih vrednosti.

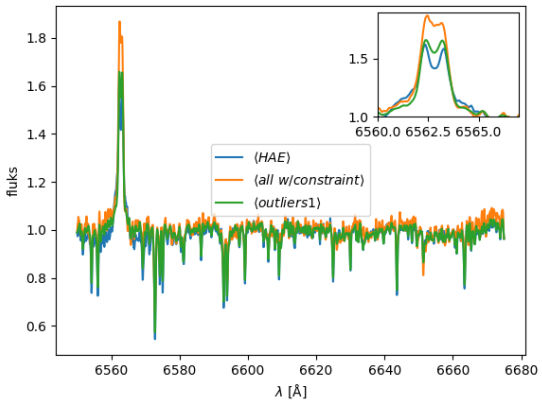
Za dober opis podatkov je potrebno vzeti dovolj največjih lastnih vrednosti oz. energije. S tem zajamemo

večino variance podatkov. Odvisnost energije od lastnih vrednosti prikazuje slika 5. V logaritemski skali lahko vidimo dva "preloma", pri $N \approx 20$ in $N \approx 100$. Podrobna analiza pokaže, da s $N \approx 7$ največjih lastnih vrednosti dobimo okrog 80% energije. To nakazuje, da je mogoče že 7 komponent PCA dovolj za opis sistema. Največ energije prispeva prva komponenta PC1, ki jo lahko najbolj verjetno pripišem efektivni temperaturi T_{eff}.



(a) (b)

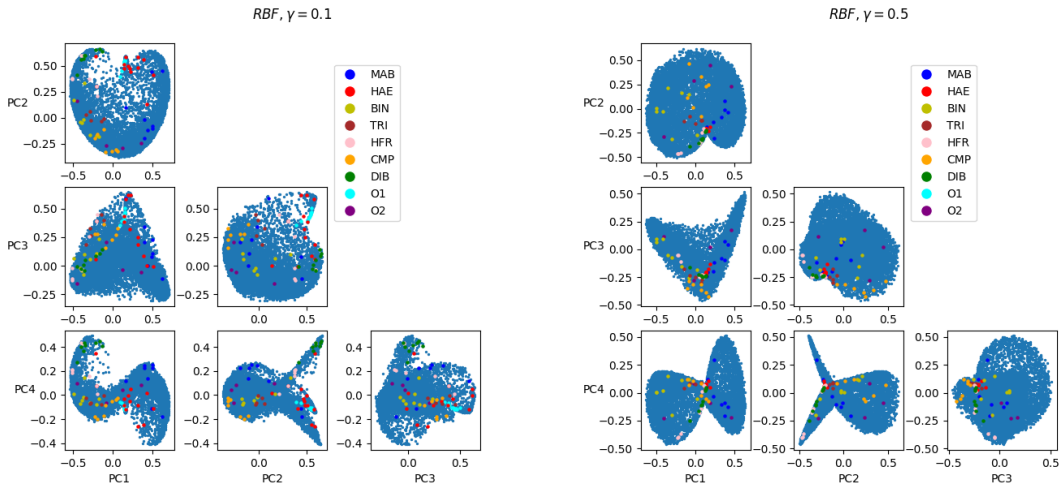
Slika 6: a) PCA kotni diagram z dodanimi učnimi seti *logg* in spektri O1, O2. b) PCA kotni diagram z dodanimi učnimi seti $[M/H]$ in spektri O1, O2.



Slika 7: Povprečni spektri v območju, kjer se nahajajo O1 (Outliers1) spektri.

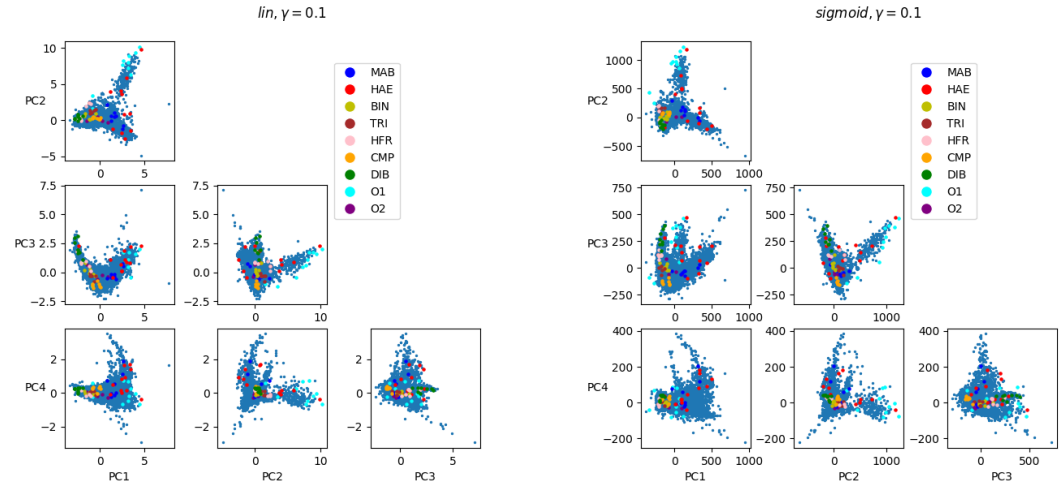
3.1 Jedrna PCA

Metodo PCA lahko razširimo z uporabo jeder. S tem v linearno metodo vpeljemo določeno mero nelinearnosti. Uporabil sem metodo $KernelPCA$ iz knjižice scikit-learn in jedra RBF, $linear(lin)$, $sigmoid$, $poly$. Pri jedrih $linear, sigmoid, poly$ nisem opazil velikih razlik. Z večanjem parametra γ so se učni spektri stiskali proti eni točki. Največje spremembe v transformacijah je podalo jedro RBF, ki je nekako po kotu ločilo učne tipe in tudi parametre v projekcijski ravnini PC1-PC2. Jedro RBF je najbolj nelinearno in bi lahko v nekaterih primeri podalo boljše razvrstitve učnih spektrov in fizikalnih parametrov od navadne PCA. V našem primeru jedrna PCA ni veliko prispevala k boljšem fizikalnem razumevanju.



(a) (b)

Slika 8: Jedrni PCA kotni diagram.



(a) (b)

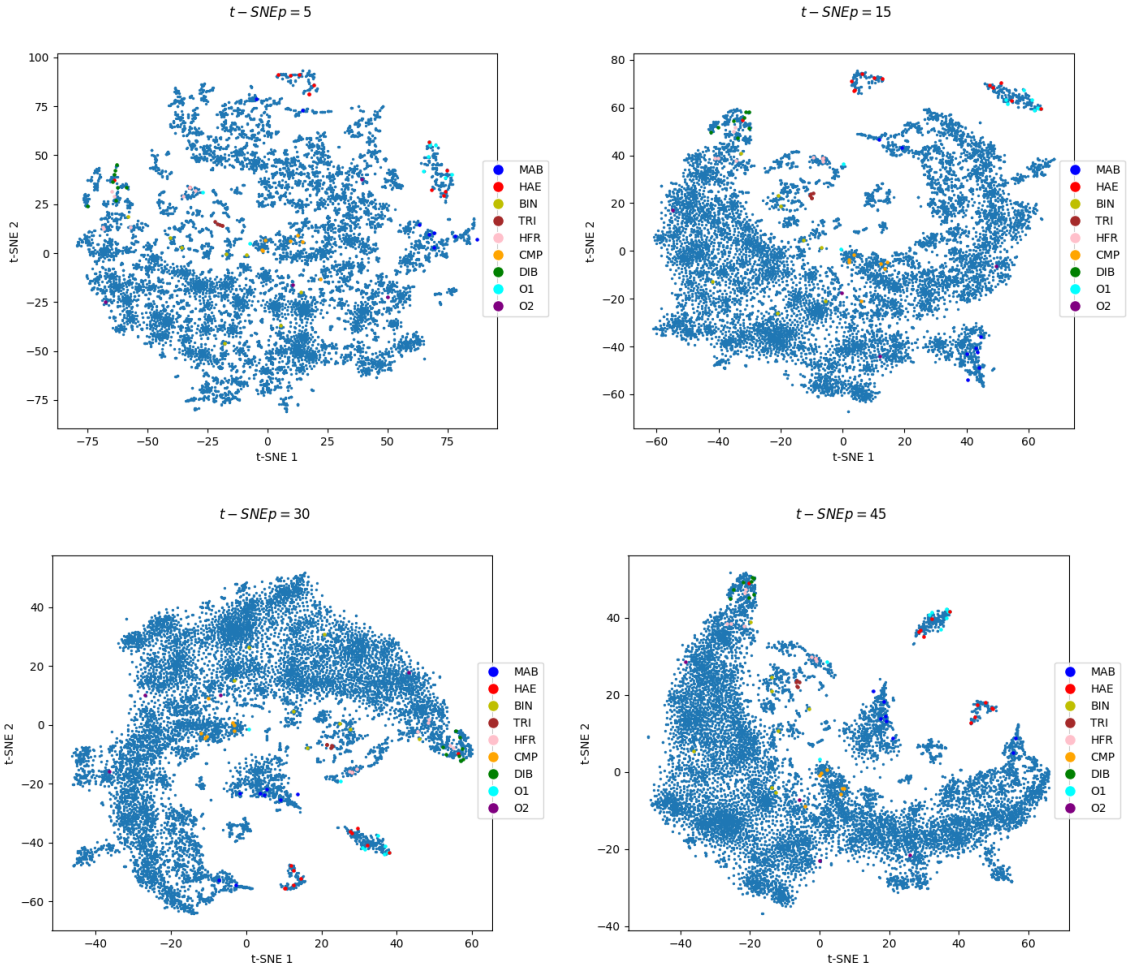
Slika 9: Jedrni PCA kotni diagram.

4 Metoda t-SNE in vizualizacija

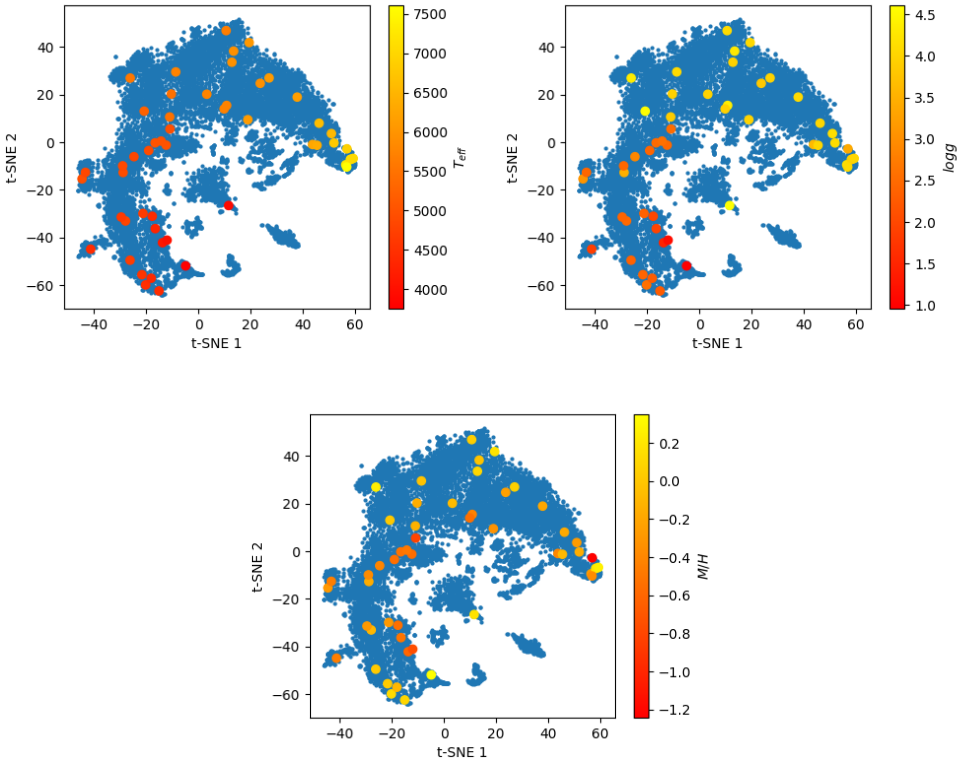
Trenutno najboljša nelinearna projekcijska metoda je t-SNE, ki projecira v 2D prostor parametrov. 2D projekcijo sestavi glede na podobnosti med točkami. Z minimizacijo t.i Kullback-Lieberjeve divergence lahko dosežemo, da v projekcijskem prostoru nastanejo gruče podobnih točk, glede na podobnost v prvotnem prostoru. Najpomembnejši parameter metode je *perplexity*, ki meri število sosedov, ki jih metoda uporablja pri računanju. Posledično lahko kontroliramo število gruč, ki nam jih metoda da. Določitev optimalnega parametra je lahko precej problematična, saj lahko gručo, ki predstavlja en fizikalni parameter razbijemo na več delov ali pa združimo več fizikalno različnih gruč v eno, ki pa nima nobene skupne fizikalne povezave. Uporabo metode pri različnih vrednostih parametra *perplexity* $\equiv p$ prikazuje slika 10. Pri vrednosti $p > 15$ nam metoda loči večino znanih učnih setov.

Projekcijo t-SNE podatkov z učnimi parametri prikazuje slika 11. Metoda lepo razvrsti spektre po temperaturi, kljub temu da je temperatura linearna kombinacija obeh spremenljivk. Temperatura se spreminja po polarnem kotu. Podobno, a ne tako zvezno odvisnosti ima tudi gravitacijski pospešek oziroma *logg*. Odvisnost kovinskosti je bolj zapletena.

V primerjavi z metodo PCA, nam t-SNE lepo loči nekatere učne sete na osamljene gruče, kar nam lahko zelo pomaga pri klasifikaciji. Metoda nam da vsaj en namig za novo gručo oziroma nov tip zvezd. Očitna izbira je eden od osamljenih večjih otokov, na katerem ni že označenih tipov. Z metodo t-SNE je mogoče tudi videti, da so spektri O1 res spektri HAE.



Slika 10: t-SNE projekcija spektrov.

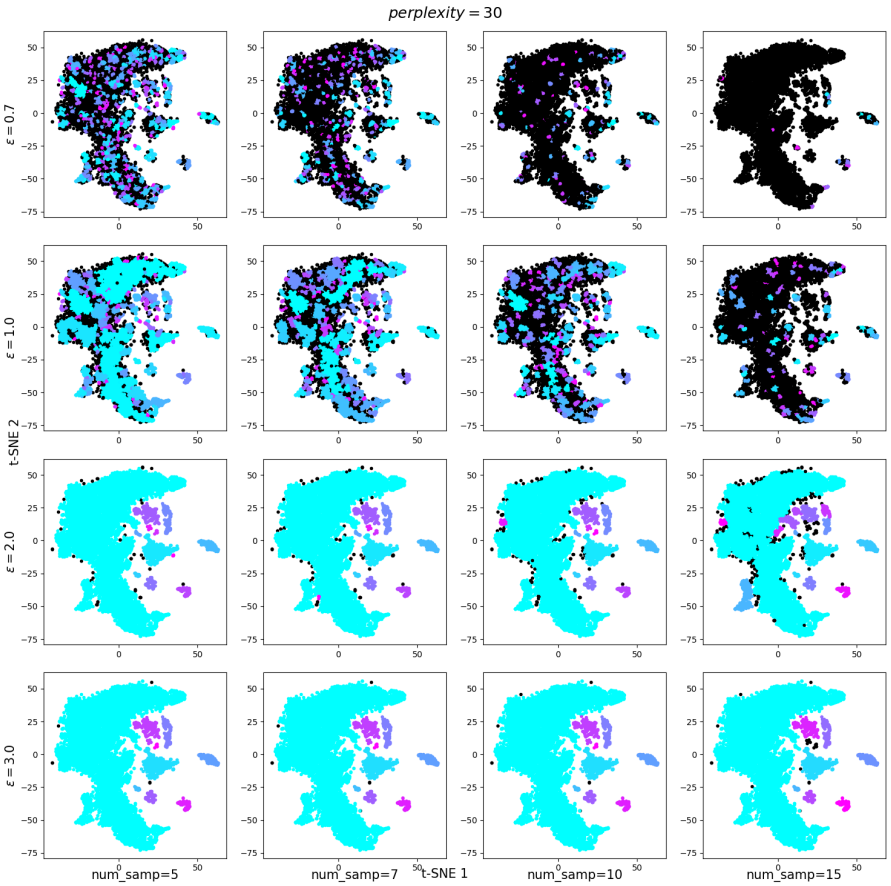


(c)

Slika 11: t-SNE projekcija spektrov, $p = 30$.

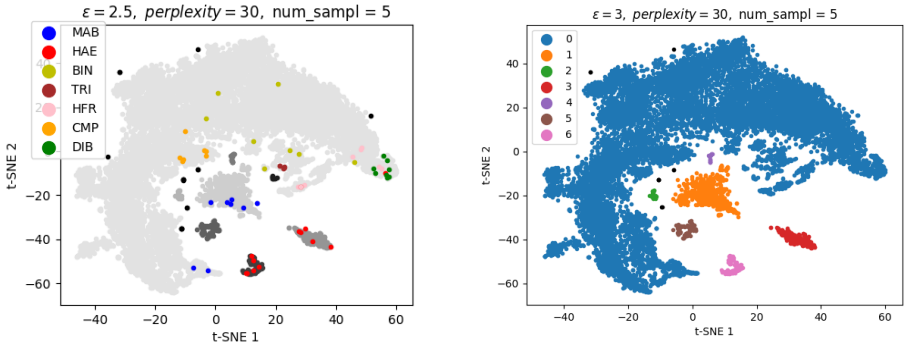
5 Iskanje gruč

Cilj naloge je poiskati kakšno novo skupino zvezd. Za iskanje gruč lahko uporabimo metodo DBSCAN iz knjižnice scikit-learn. Metoda poišče skupine točk in jih razvrsti v zaključene množice točk. Delovanje funkcije je odvisno od parametrov, najpomembnejša sta ϵ-največja dovoljena razdalja med točkama in *min_samples*-najmanjše število točk, ki lahko sestavljajo gručo. Parameter *min_samples* sem označeval z *num_sampl*. Najprej uporabimo eno od projekcijskih metod (PCA ali t-SNE) in projekcijo na ravnino. Nato lahko uporabimo metodo DBSCAN. Uporabljal sem metodo t-SNE za projekcijo. Najprej si oglejmo delovanje DBSCAN v odvisnosti od najpomembnejših parametrov (slika 12). Z naraščanjem ϵ se manjša število gruč in točk, ki niso v grupi, kar je smiselno. Z večanjem števila *num_sampl* se veča število točk, ki niso v nobeni gruči.



(a)

Slika 12: Iskanje gruč z metodo t-SNE in DBSCAN pri različnih vrednostih ϵ in *num_sampl*. S črno so narisane točke, ki ne pripadajo nobeni gruči.

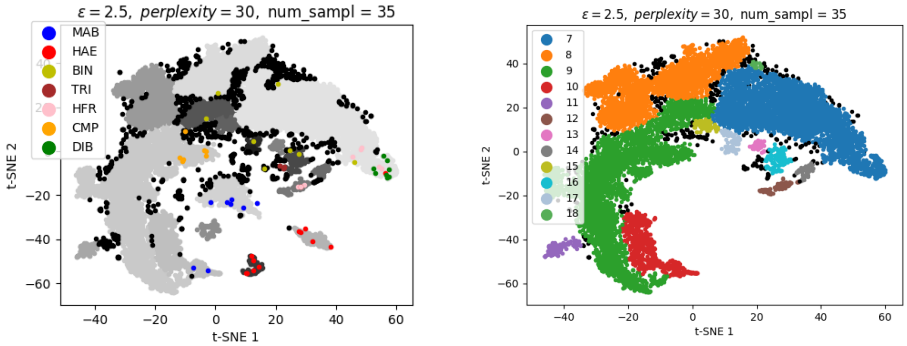


Slika 13: Iskanje gruč z metodo t-SNE in DBSCAN. S črno so narisane točke, ki ne pripadajo nobeni gruči.

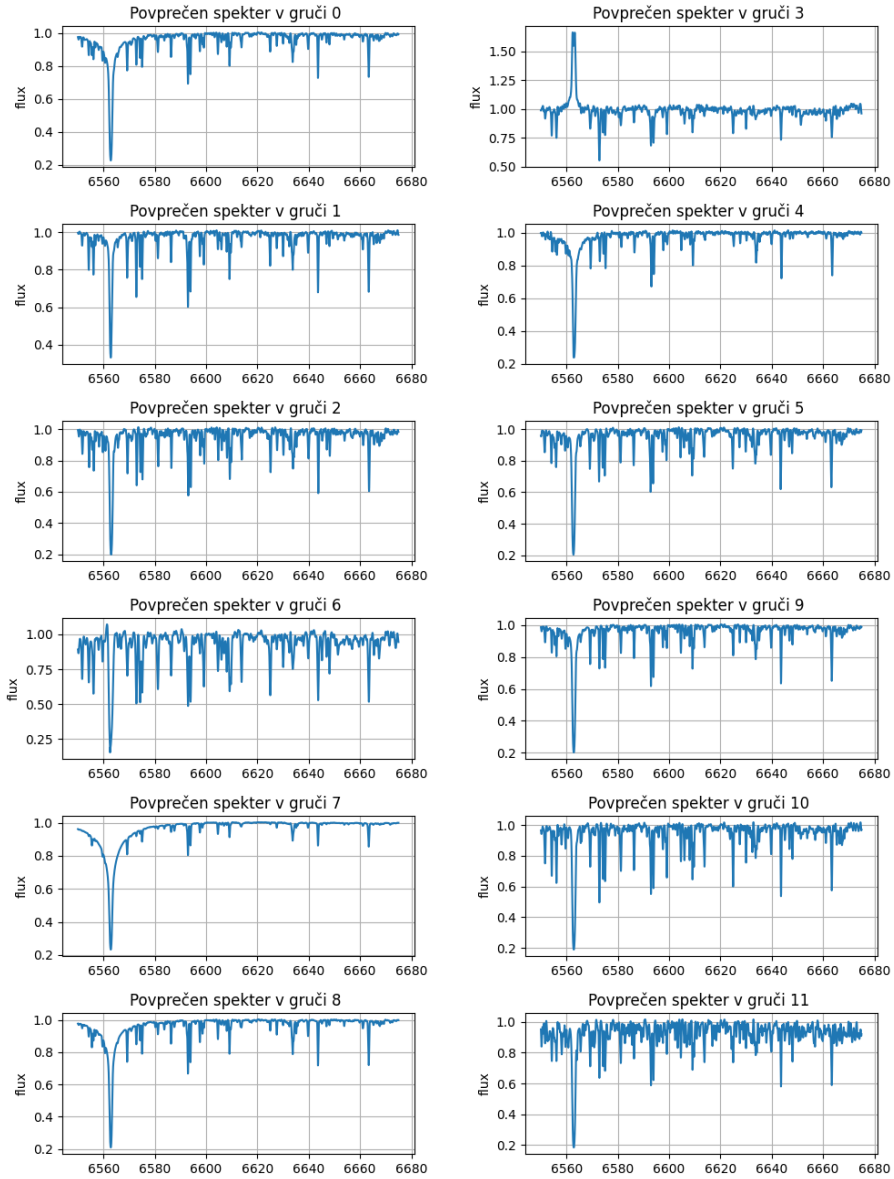
Z izbranimi parametri metode, ki so označeni na naslovih grafov lahko klasificiram naše podatke (slika 13).

Slika 13a pokaže katere gruče pripadajo že znanim zvezdam. Slika 13b pa prikazuje klasifikacijo na 7 gruč. Iz slike lahko sklepam, da bi mogoče bila gruča 5 nova, neznana skupina zvezd.

Ker je gruča 0 zelo velika, obstaja možnost da je sestavljena iz več gruč. V ta namen sem povečal število *num_sampl* in pogledal delitev gruče 0 (slika 14). S temi parametri sem gručo razdelil na 12 podgruč in za vse podgruče narisal povprečen spekter Spekter 5 (slika 15), ki je naš najboljši kandidat za novo skupino



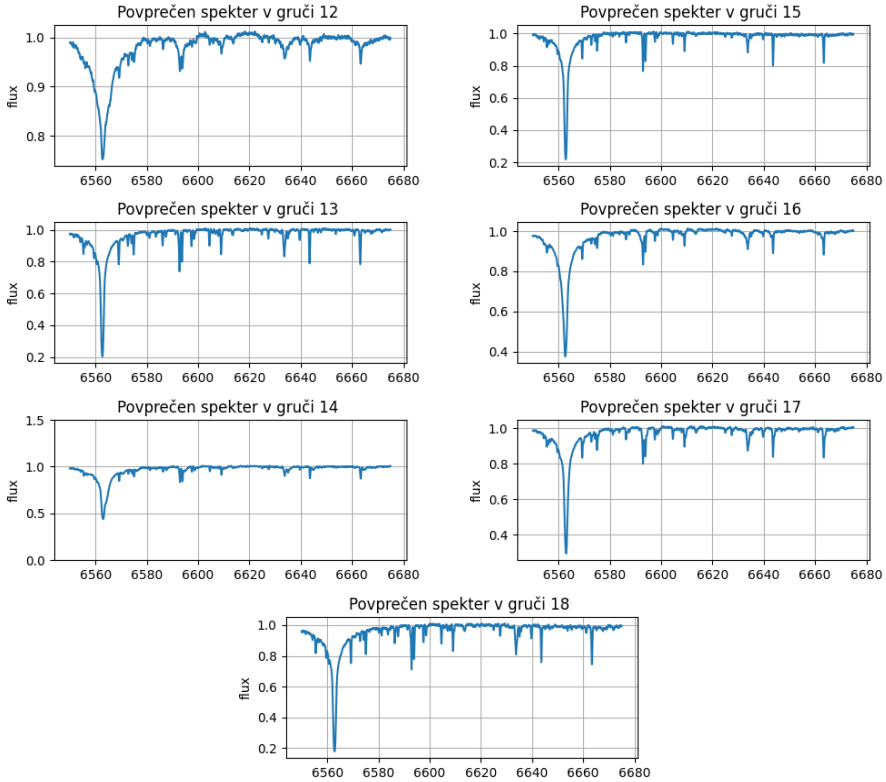
Slika 14: Iskanje gruč z metodo t-SNE in DBSCAN. S črno so narisane točke, ki ne pripadajo nobeni gruči.



Slika 15: Povprečni spektri gruč.

zvezd ima precej podoben spekter jih imajo spektri, ki pripadajo sosednjim gručam gruče 5. Vseeno se razlikuje od gruče 0 po intenziteti valovne dolžine največje absorpcije in po obliki lijaka (območje okrog

največje absorpcije). Prav tako se razlikuje po intenziteti od gruče 1. Sosednji spektri so mu podobni, a je mogoče najti razlike. V bolj podrobno analizo se ne bi spuščal saj nimam znanja na tem področju. Tudi za utemeljitev klasifikacije grupe 0 nimam dovolj znanja, zato bom to prepustil ljudem, ki se s tem ukvarjajo.



Slika 16: Povprečni spektri gruč.

```
def pca(mat, criterion):
    n = len(mat[:,0])
    mat = np.array(mat)
    u = np.mean(mat, axis=0, dtype=np.float64)
    B = np.subtract(mat, u)
    C = (1.0/(n-1.0))* B.T @ B

    eigval, eigvec = LA.eig(C)
    idx = eigval.argsort()[:: -1]
    eigval = eigval[idx]
    eigvec = eigvec[:,idx]
    W = eigvec[:,:criterion]
    B@ W
```
Slika 17: Implementirana PCA metoda.

6 Zaključek

Pri analizi spektrov zvezd smo uporabili metode PCA, t-SNE in DBSCAN, ki nam pomagajo razumeti in klasificirati različne tipe zvezd. PCA metoda je linearna in deluje bolje, ko imamo linearne zveze med količinami, v tem primeru je boljše delovala metoda t-SNE. V kombinaciji z metodo DBSCAN lahko klasificiramo različne skupine zvezd na podlagi bližine spektrov na projeciranem prostoru. Kljub temu, da nam metode zelo pomagajo grupirati zvezde, je za interpretacijo in potrditev različnih novih spektrov potrebno veliko znanja iz tega področja.