2. naloga: Klasifikacija zvezdnih spektrov

Gašper Jalen 28222055

oktober 2023

1 Uvod

Pri nalogi smo obravnali spektre zvezd, s pomočjo katerih smo želeli kvalificirati posamezne zvezde in določiti njihoove lastnosti. Osnovni koncept, ki smo ga uporabili je zmanjševanje dimenzij podatkov. Vsak zvezdni spekter namreč vsebuje 2084 točk, torej si ga lahko predstavljamo kot objekt v 2084 dimenzionalnemu prostoru. Naš cilj je zreducirati število dimenzij na npr. dve pri čemer želimo ohraniti kar največ informacije, ki nam jo spekter poda. Poleg večjega seta 10000 spektrov imamo podana še dva manjša seta podatkov, enega s klasificiranimi zvezdami in enega s podanimi fizikalnimi lastnosti zvezd (temperatura, gravitacijski pospešek na površju in kovisnkost). S pomočjo teh dveh setov lahko poizkusimo klasificirati še ostale zvezde in določit njihove lastnosti.

1.1 PCA

Sprva za redukcijo dimenzij uporabimo metodo PCA (Principal Component Analysis). Z metodo napravimo dekompozicijo podatkovnega seta v lastne vektorje in lastne vrednosti. Pri tem za željeno natančnost uporabimo željeno število lastnih vektorjev, pri čemer najprej uporabimo tiste z največjo lastno vrednostjo.

Če imamo n spektrov in ima vsak spekter p dimenzij podatki tvorijo matriko $\mathbf{X}$ dimenzij $n \times p$. Podatki so že normirani, tako da jih moramo le še centrirati pri čemer vektor $u_j = 1/n \sum_{i=1}^{n} X_{ij}$ odštejemo od stolpcov matrike $\mathbf{X}$, da dobimo matriko $\mathbf{B}$

$$\mathbf{B} = \mathbf{X} - \mathbf{h}\mathbf{u}^T, \tag{1}$$

kjer je $\mathbf{h}$ vektor enic. Nato izračunamo kovariančno matriko po definiciji

$$\mathbf{C} = \frac{1}{n-1}\mathbf{B}^*\mathbf{B} \tag{2}$$

Na koncu poiščemo še matriko lastnih vektorjev, da velja

$$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}, \tag{3}$$

pri čemer si lahko pamaga s SVD razcepom. V matriki $\mathbf{V}$ izberemo poljubno število lastnih vektorjev, ki jih zložimo v matriko $\mathbf{W}$, podatke z reducirnaim številom dimenzij pa dobimo z

$$\mathbf{T} = \mathbf{D} \cdot \mathbf{W}. \tag{4}$$

1.2 t-SNE

t-SNE metoda se uporablja za vizualizacijo visokodimenzionalnih podaktov. V našem primeru bo bolj uporabna kot metoda PCA, saj je močno linerna, kot tudi zveze med spektri zvezd in njihovimi

fizikalnimi količinami. Same metoda na tem mestu ni matematično opisana, saj je bil cilj naloge spoznavanje z metodo t-SNE in ne njena implementacija.

2 Podatkovni set

Pri reševanju naloge sem najprej prikazal nekaj značilnih spektrov, vsaek podane skupine. To sem storil tako, da sem v vsaki naslednji skupini prikazal spekter najbolj različen od prejšnje. V tem primeru sicer ne dobimo 7 spektrov, ki so najbolj različni drug od drugega, a to ni pretirano pomembno. Različnost spektrov sem sicer določil z vsoto kvadrata razlike vseh točk med posameznima spektroma. Značilni spektri so prikazani na sliki 1.

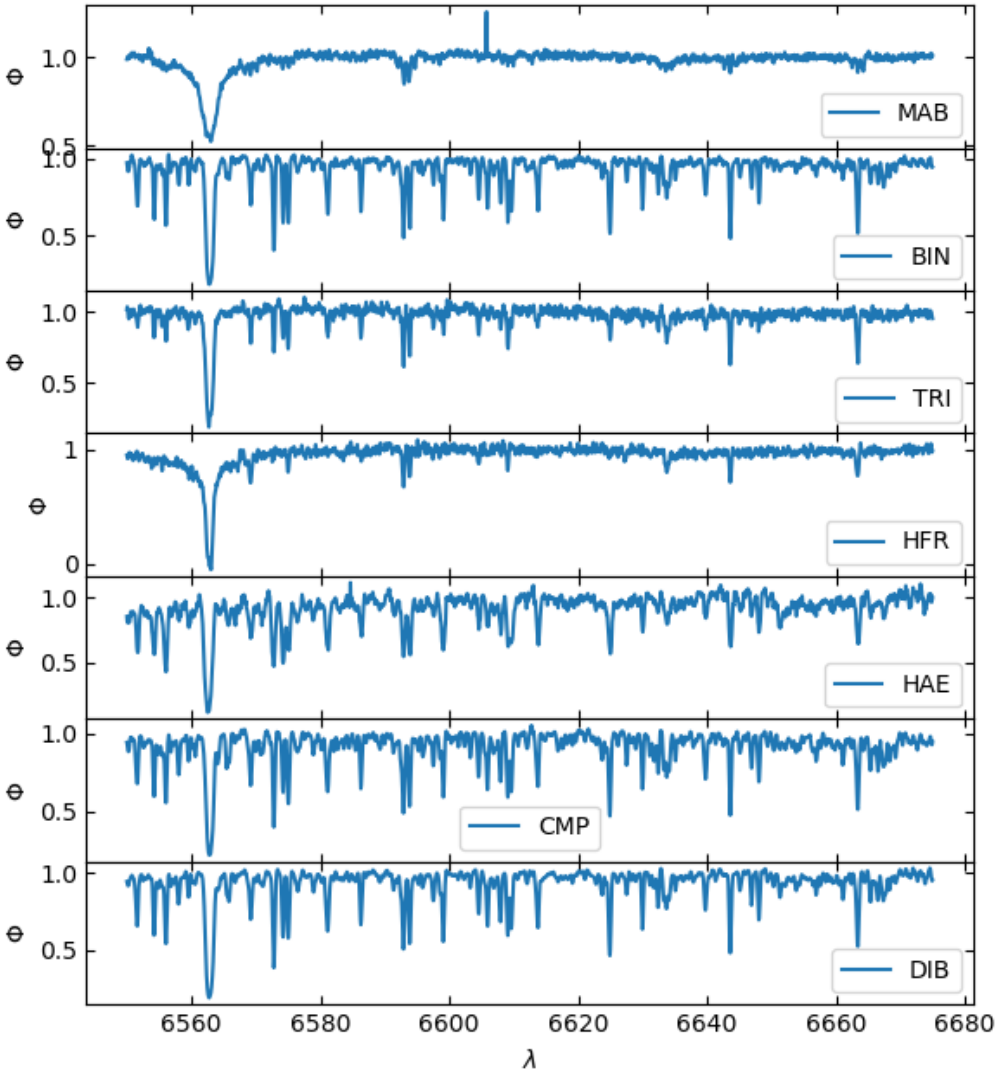

Figure 1: Primeri spektrov posameznih skupin iz klasificiranega učnega seta.

V značilnosti posameznih skupin iz učnega seta se nisem pretirano poglabljal, oznake skupin pa sem pustil take, kot so bile podane.

3 Metoda PCA

Pri metodi PCA prikažemo porazdelitve spektrov po prvih petih komponentah s pomočjo kotnega grafa na sliki 2.

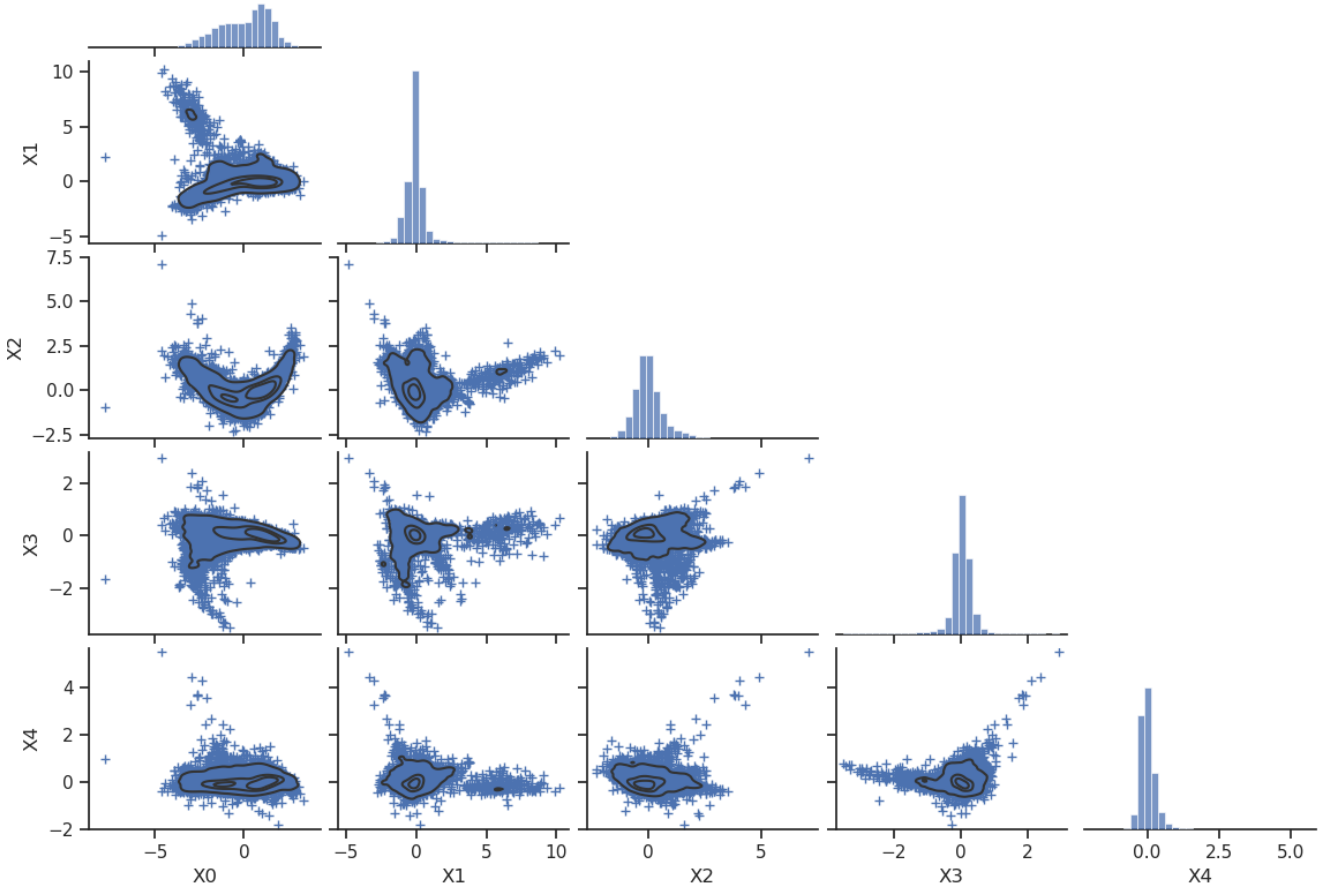

Figure 2: Porazdelitev spektrov po prvih pet komponentah metode PCA.

Pri tem nas zanima še kako dobro določeno zreducirano število dimenzij opiše originalne podatke, kar definiramo s pomočje normirane energije lastnih vektorjev

$$g_j^n = \frac{\sum_{k=1}^{j} D_{kk}}{\sum_{k=1}^{p} D_{kk}}. \tag{5}$$

če torej izberemo za aproksimacijo enako število dimenzij kot na začetku, znaša $g_p^n = 1$, kar pomeni da se ne izgubi nič informacije. Normirana energija do $j = 10$ je prikazana na sliki 3. Opazimo, da že za prvih nekaj lastnih vektorjev normirana energija doseže vrednost okoli 0.8 pri čemer začne tudi počasneje naraščati.

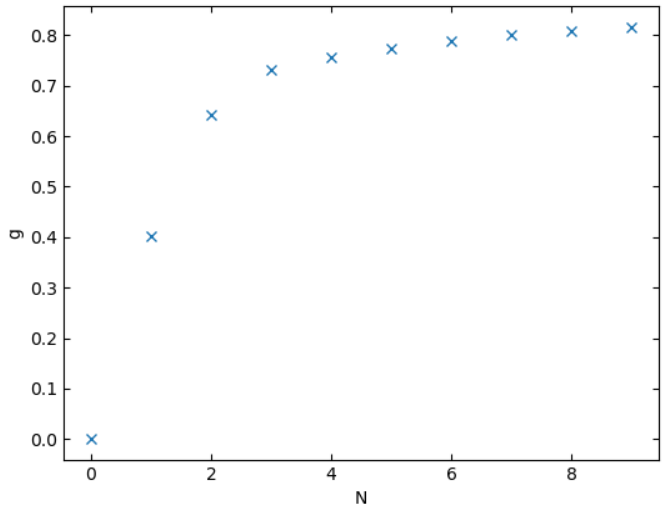

Figure 3: Porazdelitev spektrov po prvih pet komponentah metode PCA.

Zanima nas kam se pri PCA metodi preslikajo že klasificirane zvezde. Klasifikacija na kotnem grafu za prve tri komponente je prikazana na sliki 4.

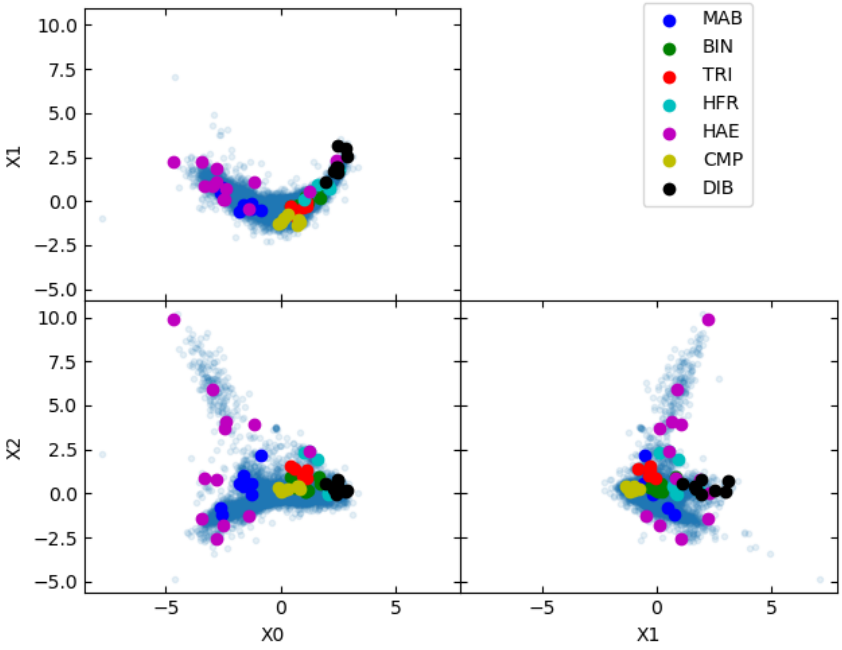

Figure 4: Porazdelitev spektrov po prvih pet komponentah metode PCA.

Klasifikacija na nek način sicer deluje, a je bi na ta način težko določili v katero skupino spada določen spekter. Območja skupin se namreč med seboj prekrivajo, ali pa med njimi ni opaziti jasne meje. Pri tem sta porazdelitvi po prvi in drugi, ter prvi in tretji komponenti približno enako razločni, porazdelitev po drugi in tretji komponenti pa ni več najbolj jasna.

Zanima nas tudi korelacija komponent s posameznimi fizikalnimi količinami, ki je prikazan na sliki 5. Največjo korelacijo opazimo med prvo komponento in temperaturo, pri čemer temperatura narašča z vrednostjo komponente. Pri temperaturi in ostalih dve komponentah, je sicer še mogoče prepoznati neko značilno smer v katero so podatki bolj razpršeni, prav tako pri korelaciji med g in prvima dvema komponentama, medtem ko za ostale primere skupki točk izgledajo naključni.

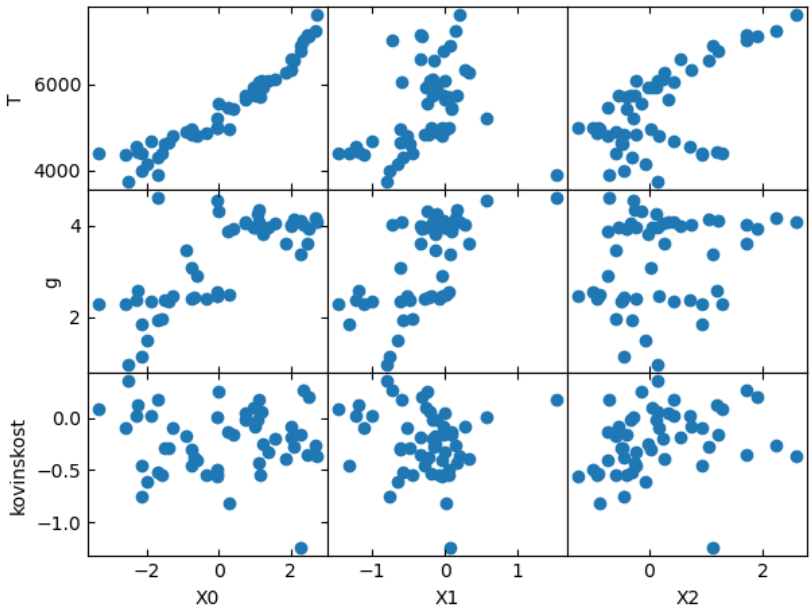

Figure 5: Korelacija med fizikalnimi količinami in novimi komponentami dobljenimi s PCA metodo. Prikazani so rezultati za prve tri komponente.

3.1 Kernel PCA

Dodatno preverimo še delovanje metode kernel PCA, ki za razliko od metode PCA ni popolnoma linearna. Pri metodi sem testiral kako uspešno se ločijo že znani razredi, kar sem prikazal na kotnem grafu za prve tri komponente. Pri tem sta se kot nekoliko bolj uspešni izkazali jedri cosine in poly (polinom pete stopnje), kar je prikazano na slikah 6 in 7. Pri ostalih jedrih so bili rezultati precej podobni običajni PCA metodi in jih nisem prikazal.

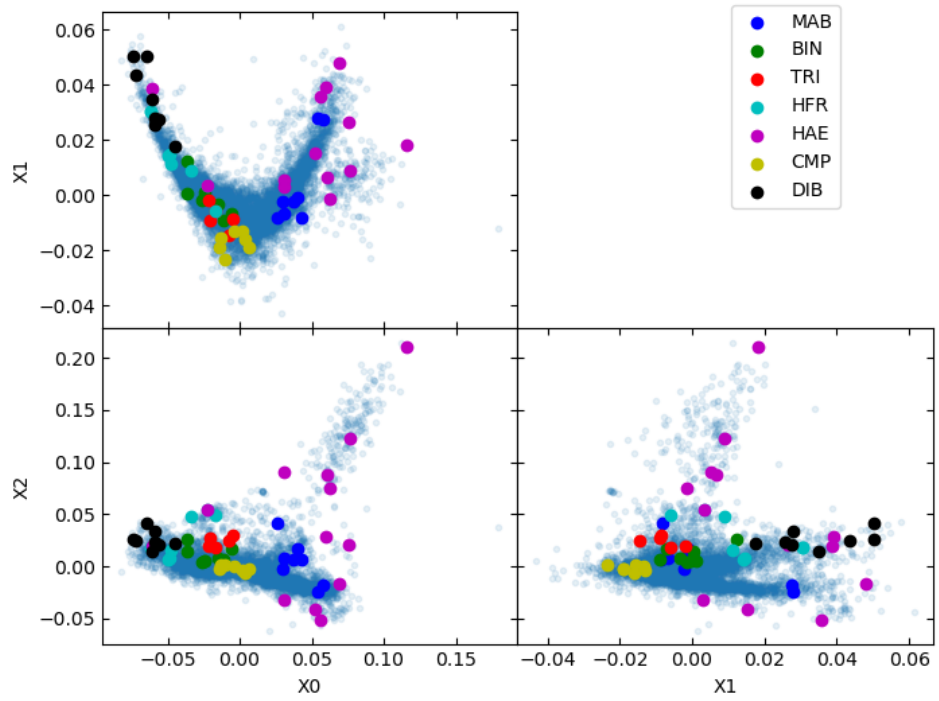

Figure 6: Porazdelitev spektrov po prvih pet komponentah metode PCA kernel (jedro cosine). Prikazani so rezultati za prve tri komponente.

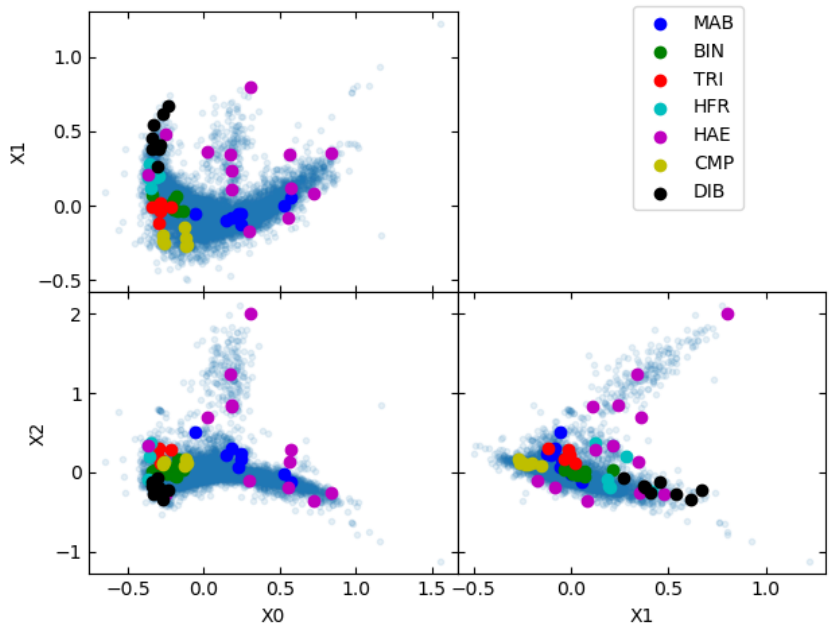

Figure 7: Porazdelitev spektrov po prvih pet komponentah metode PCA kernel (jedro poly5 - polinom pete stopnje). Prikazani so rezultati za prve tri komponente.

4 t-SNE

V drugem delu naloge opravimo redukcijo dimenzij z metodo t-SNE pri čemer obravnavamo le redukcijo na dve dimenziji. Pri tej metodi je potrebno spreminjati parameter perplexity (p), ki določa kako metoda grupira posamezne spektre. Rezultati porazdelitve spektrov za različne vrednosti parametra so prikazane na sliki 8.

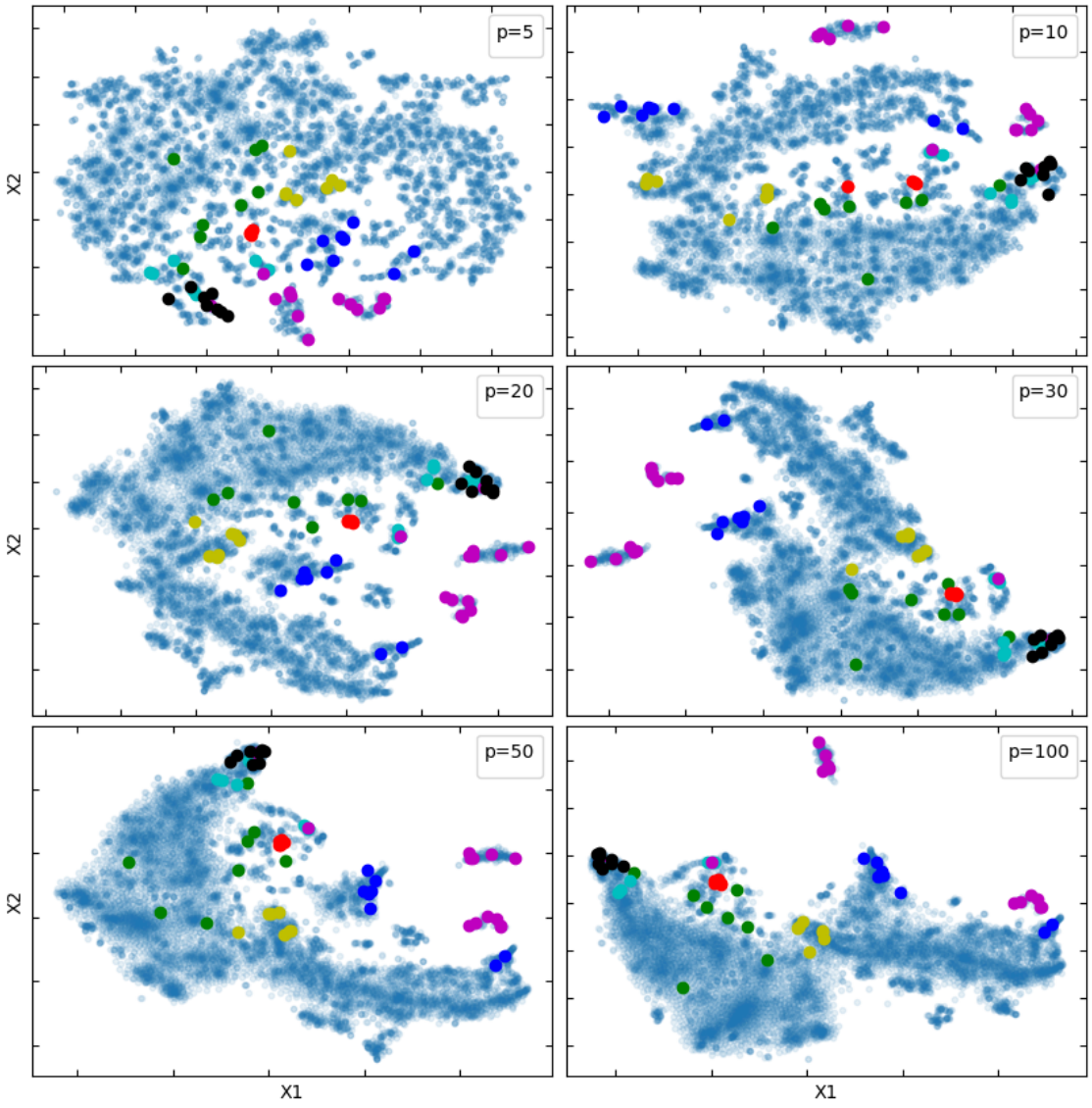

Figure 8: Porazdelitev spektrov po prvih dveh komponentah metode t-SNE za različne vrednosti parametra p. Dodatno so označene že znane skupine zvezd.

Opazimo, da metoda za $p = 5$ in $p = 10$ spetkrov ne loči najbolje, medtem ko so rezultati za višje vrednosti parametra boljši. Kljub temu pa nobena vrednost posebno ne izstopa po tem da bistveno bolje razloči med posameznimi skupinami, zato sem v nadaljevanju uporabljal parameter $p = 20$. Pri tem nekatere skupine zvezd izstopajo precej bolj kot druge. Posebej značilna je skupina HAE (vijolična barva), ki ima v spektru značilno emisijsko črto in jo metoda za večino podanih članov skupine jasno loči od ostalih spektrov. Zanimivo se skupina sama pri vseh vrednosti p razdeli na dva otoka.

Poleg klasifikacije nas ponovno zanima korelacija s fizikalnimi količinami. Te so na sliki 9 prikazane z barvo točk na ravnini prvih dveh komponent. Ponovno opazimo največ reda pri temperaturi, kjer ta narašča po polkrožni obliki od zgoraj navzdol. Podoben a manj jasen vzorec opazimo pri pospešku, medtem ko izgleda kovinskost zvezd naključno porazdeljena po obeh koordinatah.

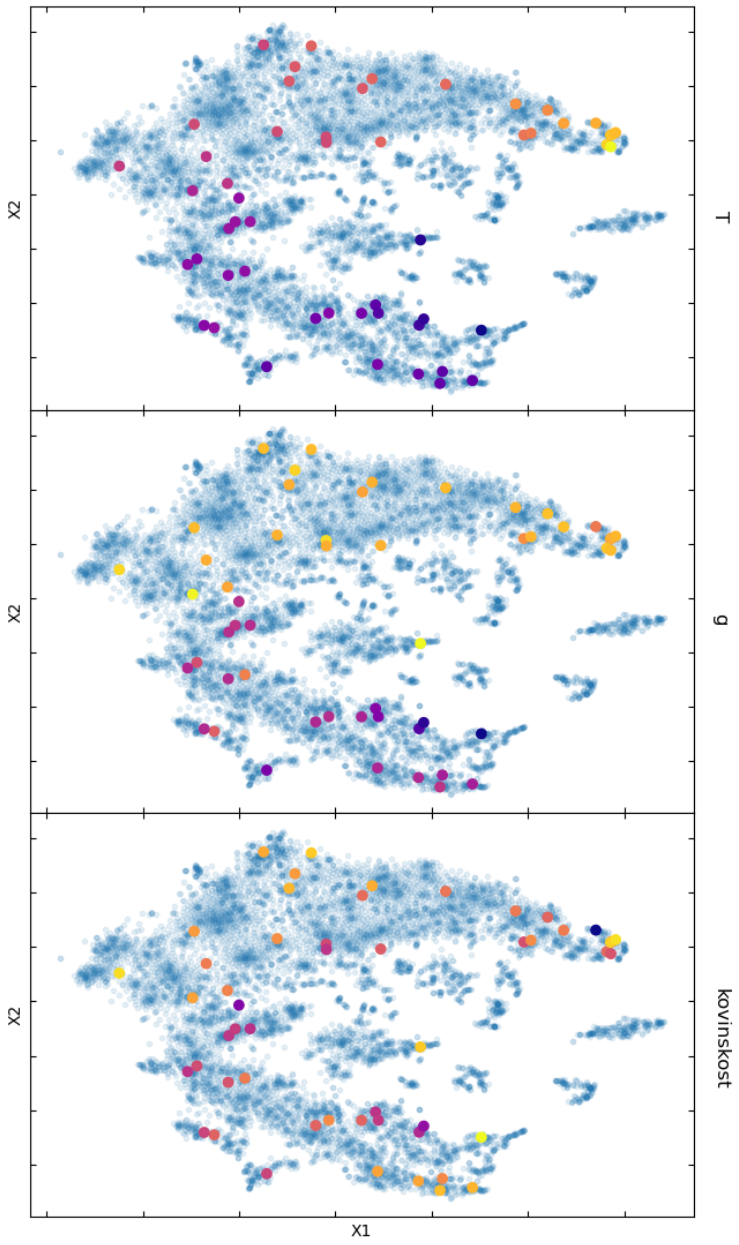

Figure 9: Fizikalne količine zvezd prikazane z barvnimi točkami na porazdelitvi zvezd po prvih dveh komponentah metode t-SNE. Količine so normirane zato skala ni prikazana.

V duhu prejšnje naloge izkoristimo dejstvo, da je temperatura kolerirana z obema koordinatama in poiskusimo s fitom določiti temperaturo vseh zvezd. Za to uporabimo proces GPR z jedrom Matern. Z nekaj prilagajanji parametra ν (za končni fit je uporabljena vrednost $\nu = 2.5$) dobimo podatke o temperaturi za vse spektre, kar je prikazano na sliki 14.

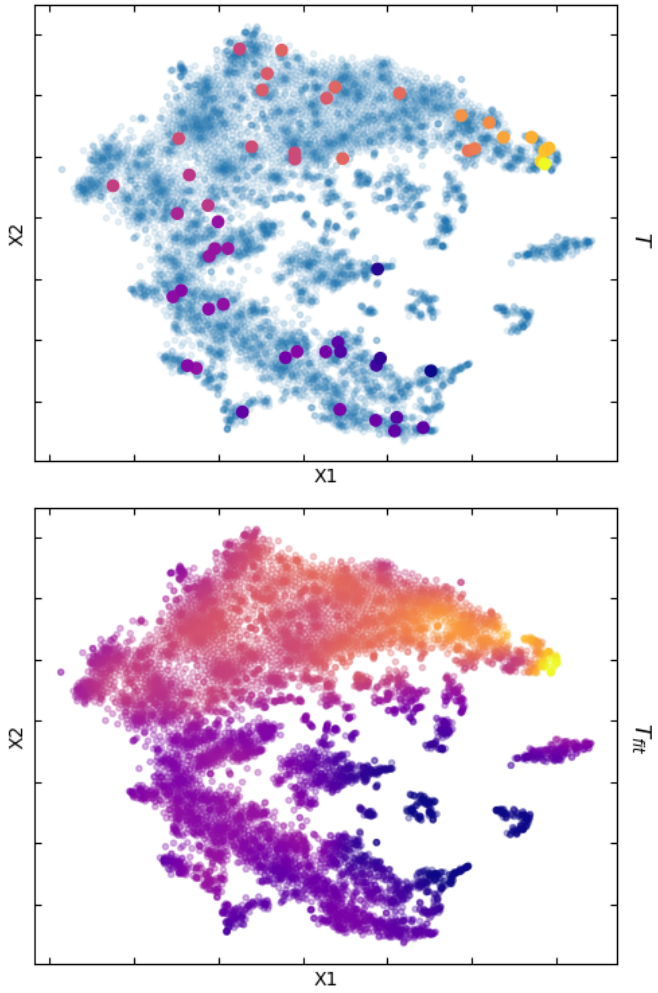

Figure 10: Temperatura zvezd prikazana z barvnimi točkami na porazdelitvi po prvih dveh komponentah metode t-SNE za znane spektre (zgoraj) in za vse spektre (spodaj).

5 DBSCAN

Na koncu naloge preizkusimo še avtomatsko klasifikacijo podatko z metodo DBSCAN. Metodi podamo dva parametra in sicer ε, ki vpliva na oddaljenost točk znotraj ene grupe, in minimalno število vzorcev v posamezni grupi (min_samp). Rezultati klasifikacije za različne vrednosti parametra so prikazani z različnimi barvami na sliki. Za neko smiselno klasifikacijo uporabimo vrednosti $\varepsilon = 1.8$ in min_samp $=$ 15, pri katerih z gručami pokrijemo večino točk, gruče same pa niso pretirano velike.

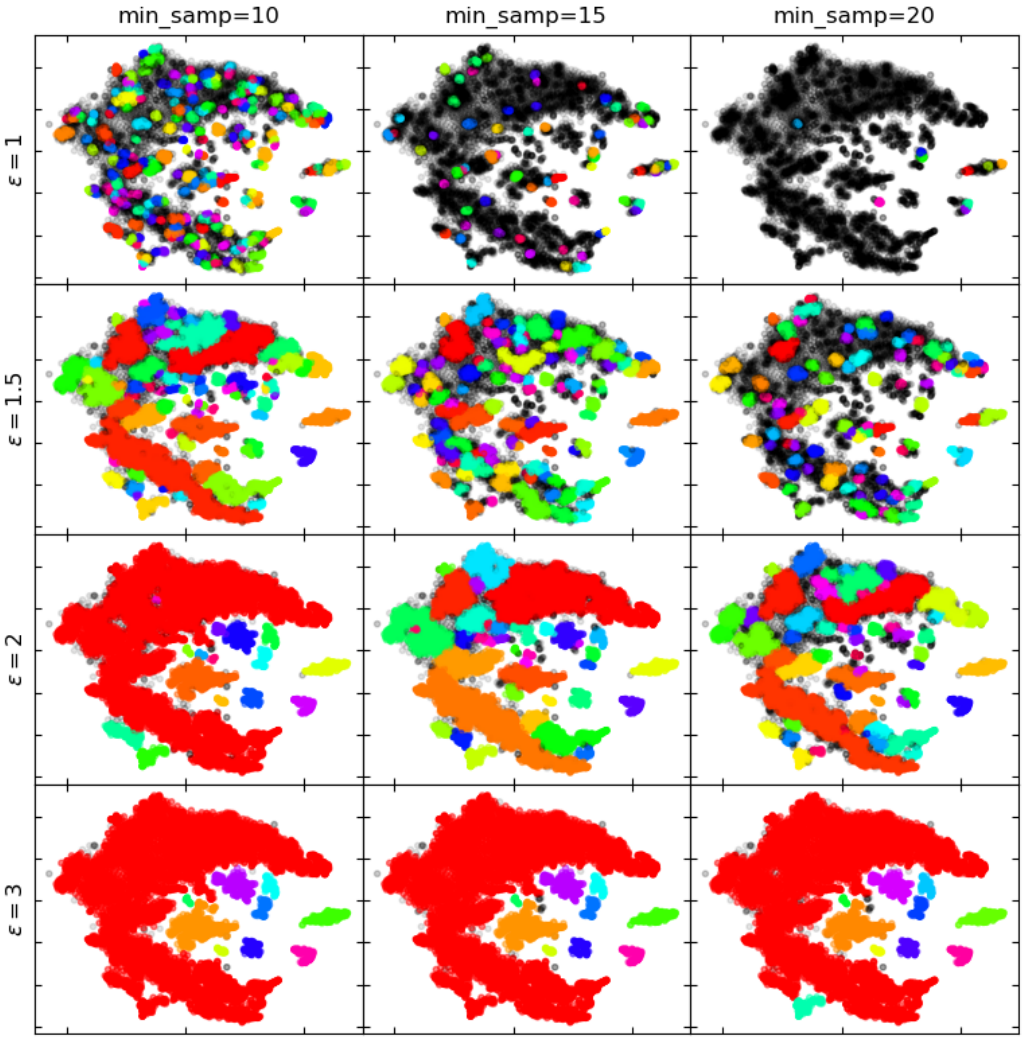

Figure 11: Klasifikacija točk z metodo DBSCAN za različne parametre ε in min_samp. Posamezne gruče so označene z različnimi barvami s črno pa je prikazan šum.

Glede na klasifikacijo z metodo DBSCAN želel v podane skupine v učnem setu razvrstiti še ostale spektre. To sem preprosto storil tako, da sem pogledal v katerih gručah leži največ točk posamezne skupine in vse zvezde iz teh gruč pripisal posamezni skupini. Na tak način sicer večina zvezd ostane neklasificiranih, so pa zato klasificirane zvezde z večjo verjetnostjo pravilno klasificirane. Dodatno sem preveril še eno značilno gručo, ki jo je DBSCAN pri večini parametrov ločil od ostalih in sicer sem jo označil z GA. Na sliki 12 sem prikazal kako spektri iz te gruče odstopajo od nekega tipičnega povprečnega spektra celotnega seta. Izkaže se da imajo spektri iz tovrstne gruče globlje absorbcijske črte.

Na sliki 13 je prikazana klasifikacija spektrov po skupinah podanih v učnem setu z dodano skupino z globokimi absorbcijskimi črtami. Na podoben način bi verejtno lahko obravnavali vse gruče, ki jih DBSCAN razloči (pri čemer bi si bili določeni spektri med seboj zelo podobni), vendar pa bi zaradi malo znanja o spektrih zvezd na ta način težko prišel do jasnih zaključkov o tipih zvezd. Prav tako mi na tem mestu ni bilo popolnoma jasno ali lahko s klasifikacijo v učnem setu predstavimo vse spektre ali pa del spektrov tvori drugi nepoznane skupine.

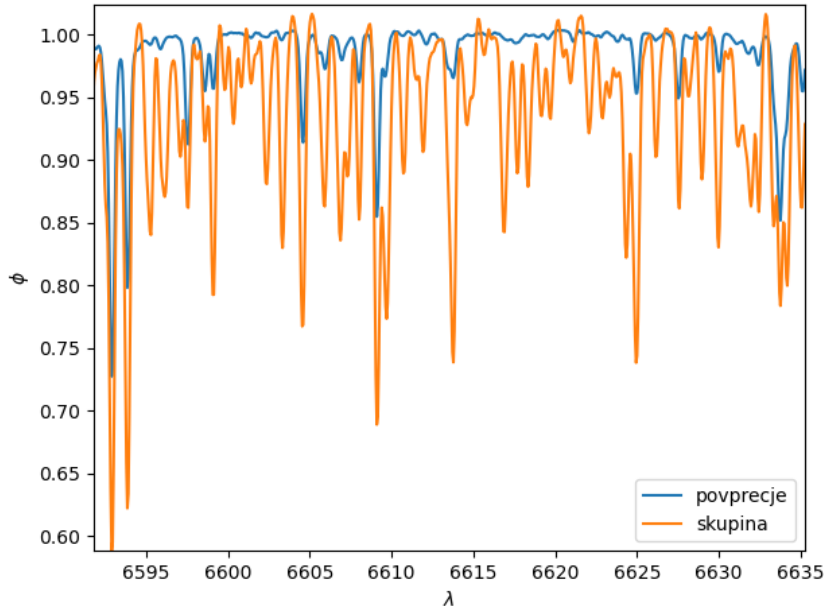

Figure 12: Del tipičnega povprečnega spektra in spektra iz izolirane gruče.

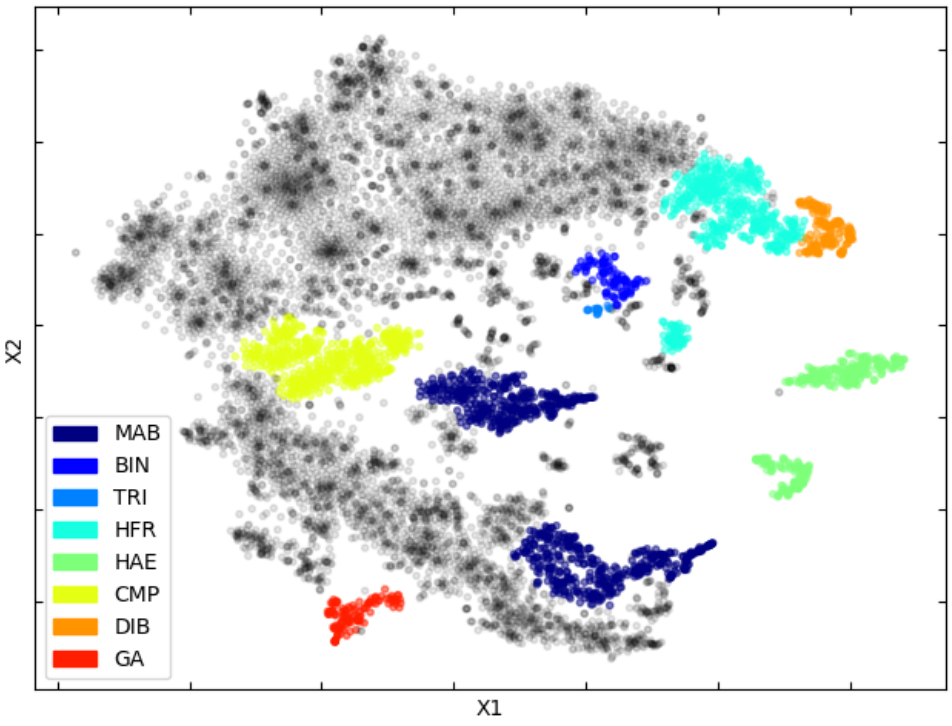

Figure 13: Klasificarni spektri. Pobarvani so spektri, ki spadajo v skupine iz učnega seta in še spektri, iz skupine GA.

6 Dodatek

```python
def PCA(DATA, n):

    # center each dimension
    average_val = np.sum(DATA, axis=0)/DATA.shape[0]
    DATA -= average_val[np.newaxis, :]

    # covariance matrix
    cov = 1/DATA.shape[0] * np.matmul(DATA.transpose(), DATA)

    U, S, Vh = np.linalg.svd(DATA, full_matrices=True)

    sigma_square = S**2

    energ = np.sum(sigma_square)

    # vrne array ucinkovtiosti
    eff = [np.sum(sigma_square[:i])/energ for i in range(10)]

    # vrne le vrednost
    #eff = np.sum(sigma_square[:n])/energ

    # matrix from eigenvector (lines of Vh)
    W = Vh[:n].transpose()

    return W, eff
```

Figure 14: Moja implemenacija PCA metode.