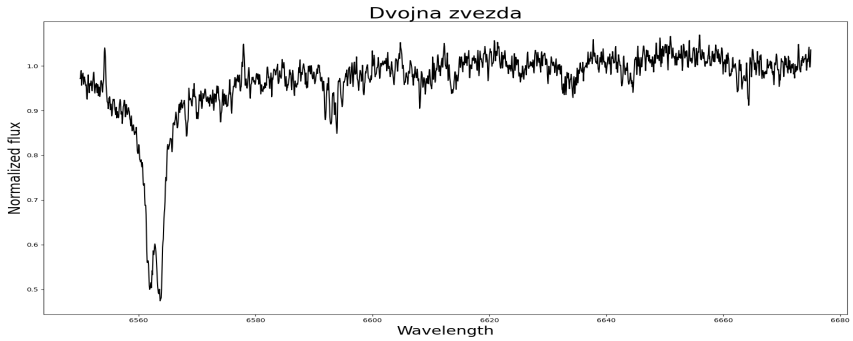
Praktikum strojnega učenja v fiziki: 2. naloga - Klasifikacija zvezdnih spektrov s PCA in t-SNE

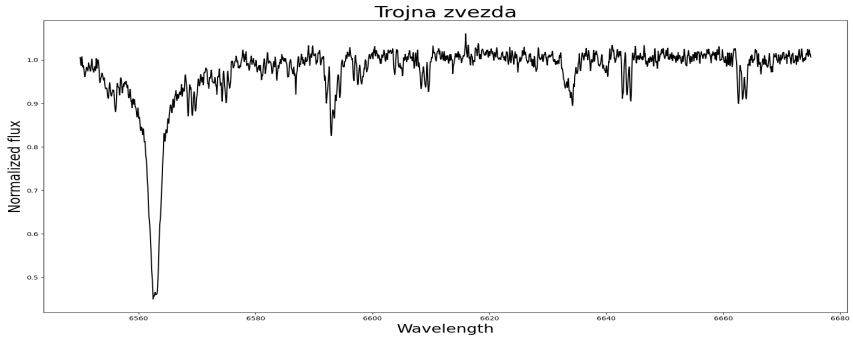
Matic Debeljak

6. november 2020

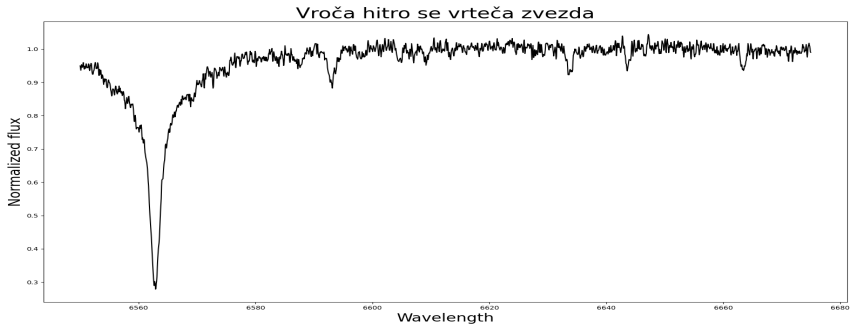
1 Podatkovni set

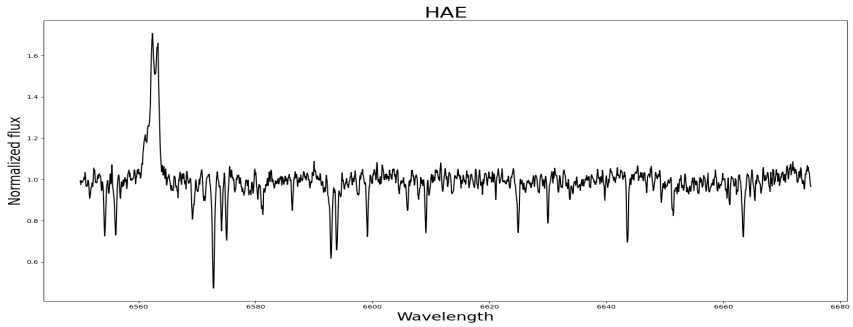
Podatkovni set za to vajo je 10000 spektrov širokih 125 Å, ki so jih izmerili pri projketu GALAH. Za lažjo predstavo bom narisal nekaj od teh spektrov.

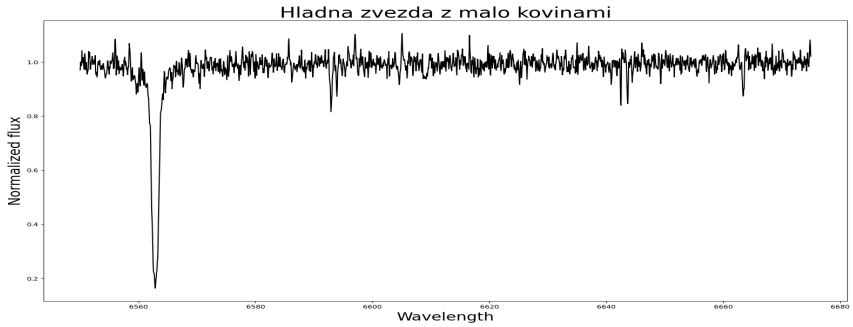


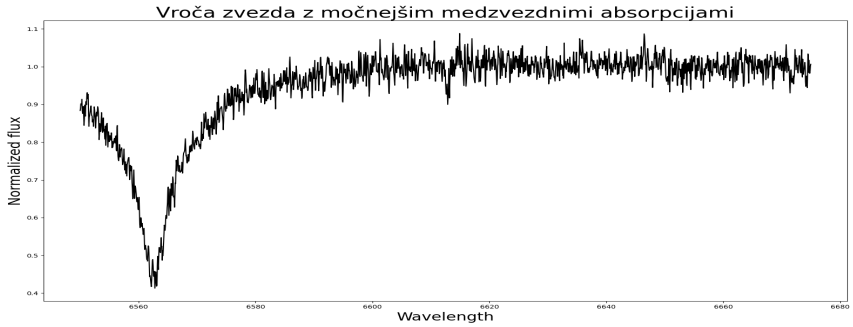


Kot vidimo na slikah so si spektri med sabo doakj podobni a vseno lahko opazimo med njimi razlike, glede na te razlike so zvezde razdeljene v različne skupine. Cilj te naloge je zvezde ločiti glede na njihov spekter in jih razdeliti v logične skupine, ter to vizualzirati.


Vroča hitro se vrteča zvezda
Normalized flux
Wavelength

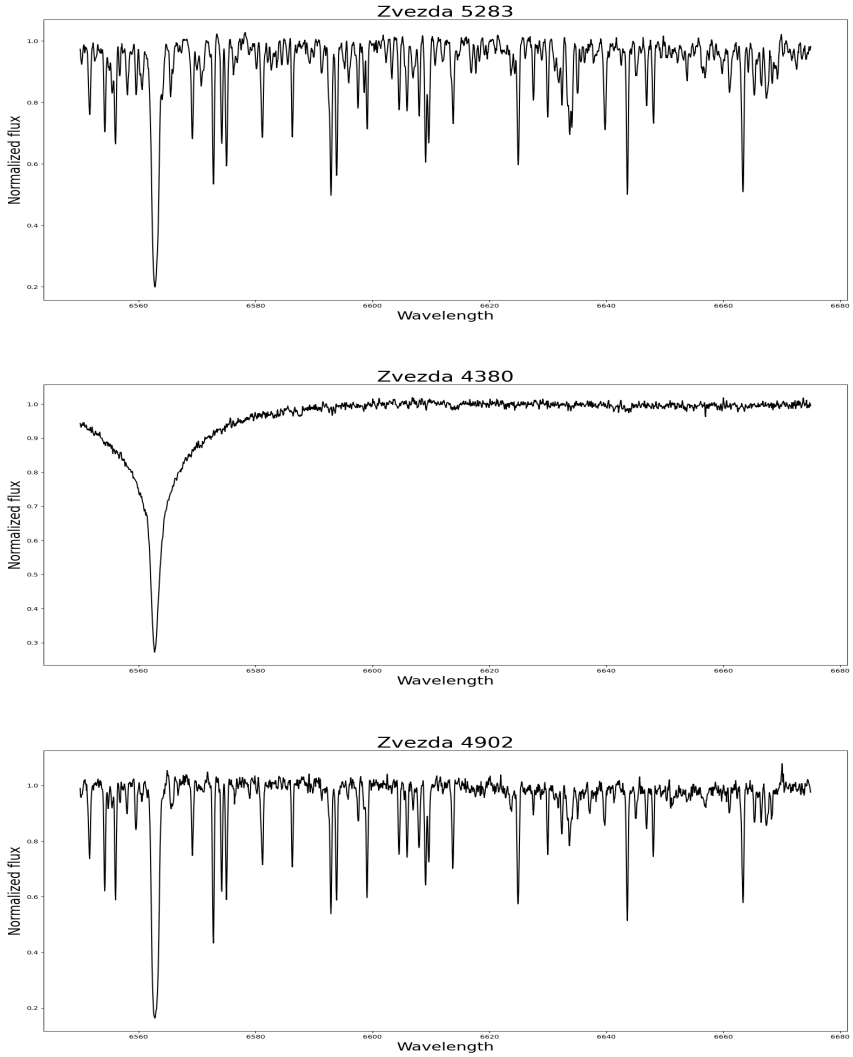

HAE
Normalized flux
Wavelength


Hladna zvezda z malo kovinami
Normalized flux
Wavelength


Vroča zvezda z močnejšim medzvezdnimi absorpcijami
Normalized flux
Wavelength

2 Zanimivi spektri

Nekateri spektri, pa se popolnoma razlikujeo od ostalih in se jih neda uvstriti v nobeno od glavnih skupin. Nekaj teh je predstavljenih na spodnjih slikah.

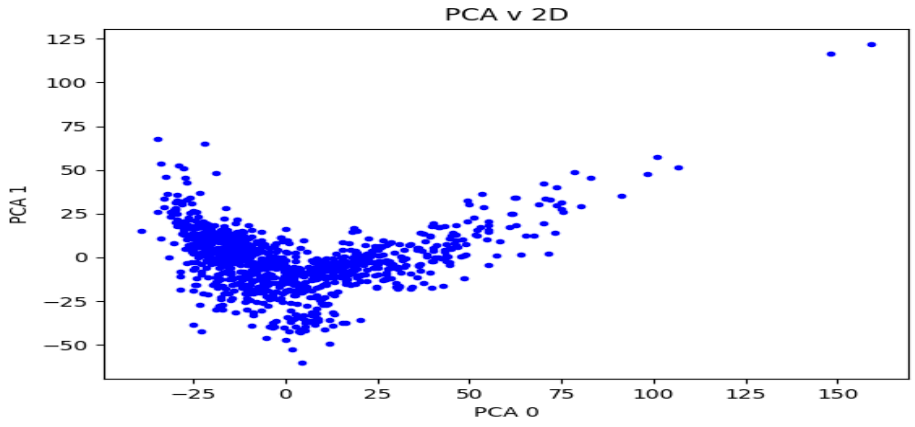


3 Zmanjševanje dimenzij

Naš podatkovni set je sestavljen it 10000 spektrov, vsak od spektrov pa je opisan p = 2084 spremenljivkami. Naš cilj je vsakega od spektrov opisat s l ($l < p$) spremeljivkami. Želimo si, da l = 2 ali l = 3, da bi lahko spektre prikazli na 2D ali 3D grafu, hkrati pa pri tem izgubili čimanj podatkov. Prav tako si želimo, da bi naše spremenljivke predstavljale neko fizikalno količino na primer temperaturo zvezde. Za tako redukcijo, bi potrebovali izjemno zapleten algoritem. V praksi pa nove spremenljivke predstavljajo neko neznano kombinacijo fizikalnih količin. Pri tej nalogi bomo za zmanjševanje dimenzij uporabili dve zelo pogosto uporabljeni metodi in sicer PCA in t-SNE.

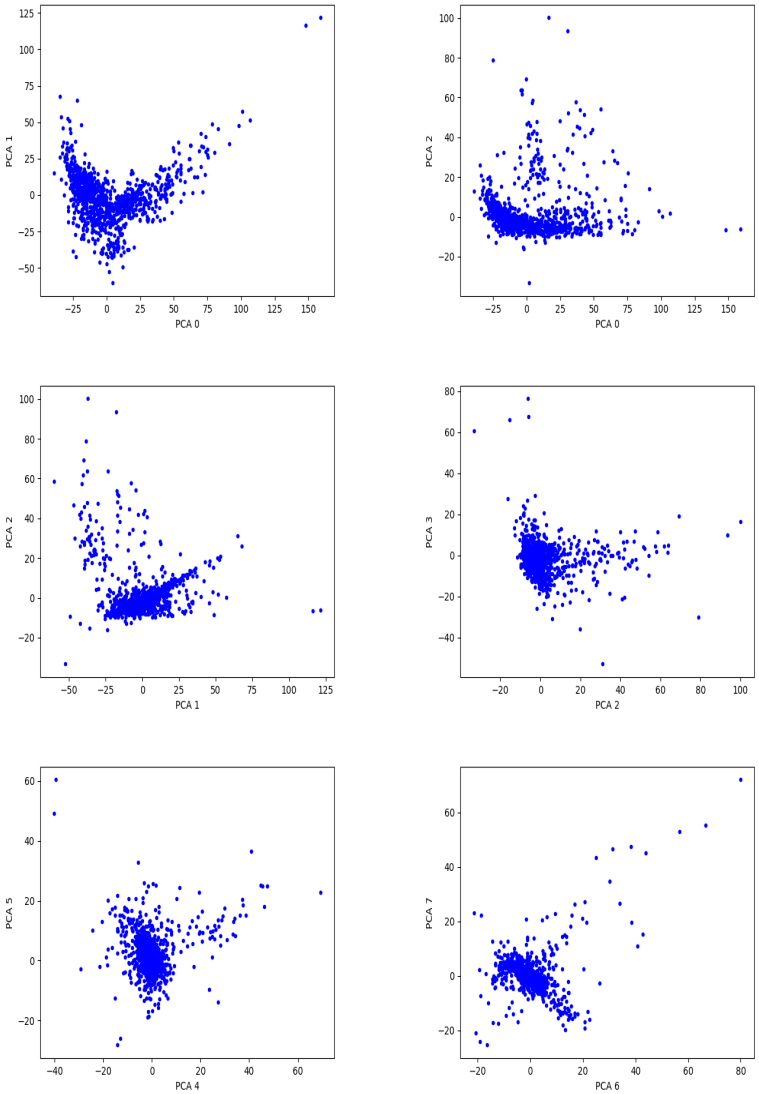
3.1 PCA

Metoda glavnih komponent (Principal component analysis – PCA) je ena od osnovnih metod zmanjševanja dimenzij. S to metodo naredimo dekompozicijo našega podatkovnega seta v lastne vektorje in lastne vrednosti. Lastne vektorje lahko rangiramo od najbolj zastopanih v podatkovnem setu do najmanj zastopanih v podatkovnem setu. Podatkovni set nato lahko predstavimo z linearno kombinacijo nekaj najbolj zastopanih lastnih vektorjev.
Metodo PCA se pogosto uporablja saj je linearna in hitra, njena pomankljivost pa je, da ni vedno lahko določiti na koliko spremenljivk naj metoda reducira podatkovni set, ter da je ta številka ponavadi večja od 3, zato je podatke težko vizualno predstaviti. Pa poskusimo najprej s projekcijo v dve dimenziji (slika 1).



Slika 1: Metoda PCA v dve dimezije.

Opazimo, da je na sliki ena velika gruča zvezd in nekaj zvezd ki odstopajo. Ne opazimo pa rezultata, ki ga pričakujemo oziroma več različnih manjših skupin.
Kot sem omenil prej je ena od težav PCA, da je težko določiti število dimenzij na katere bomo zreducirali naš podatkovni set, sam sem se po nekaj različnih poskusih odločil za 9 (slika 2).



Slika 2: Metoda PCA v devet dimezij.

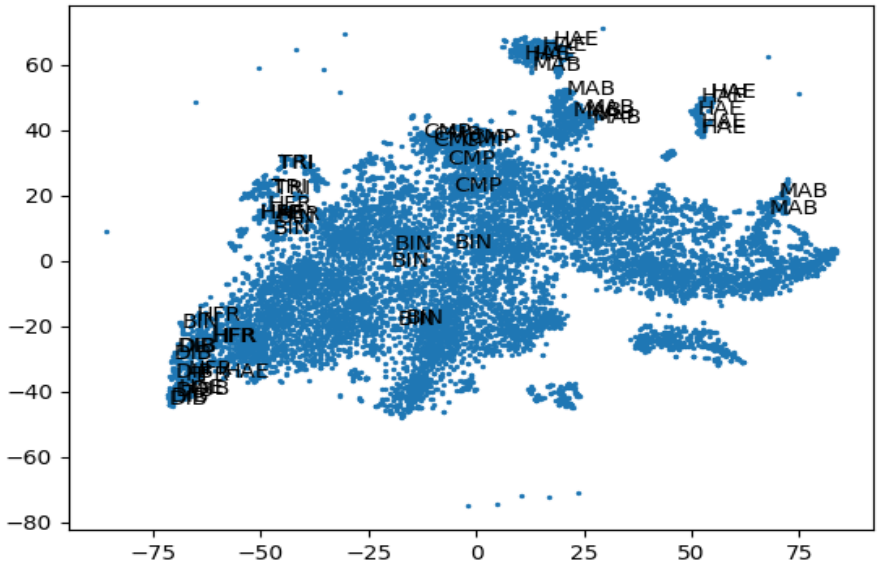
Kljub temu, da dobimo s projekcijo v več dimenzij veliko več podatkov,

ki imajo vrjetno tudi nekakšen fizikalni pomen, si je iz grafov, ki sem jih dobil zelo težko kaj predstavljati, zato se odločimo za metodo t-SNE, ki naj bi bolje prikazla gruče posameznih zvezd.

3.2 t-SNE

t-SNE (t-distributed stochastic neighbour embedding) je močno nelinearna metoda. V glavnem se uporablja za vizualizacijo visokodimenzionalnih podatkov, mi pa jo bomo uporabili kot algoritem za klasifikacijo zvezd. Ceprav lahko število dimenzij zreducira poljubno, bomo v naših primerih set podatkov vedno zreducirali na dve dimenziji.

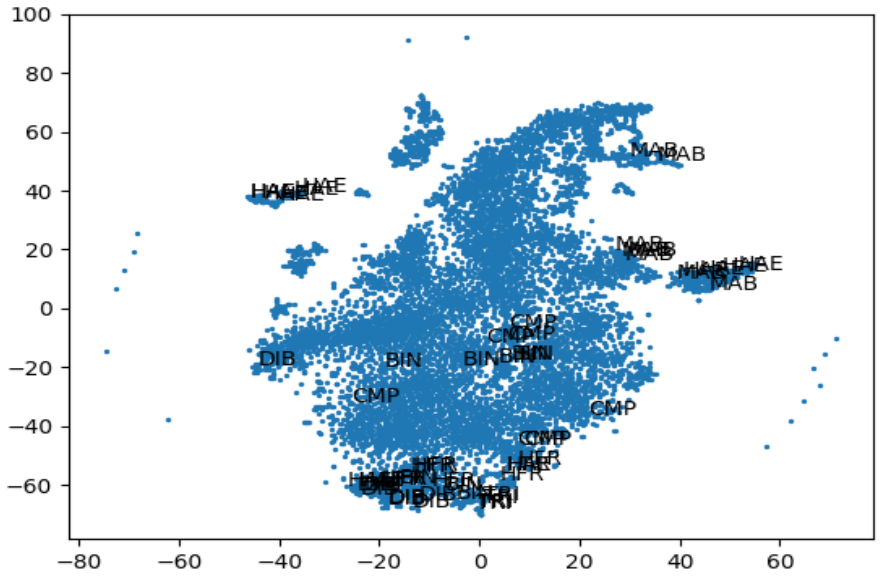
Kljub temu, da je metoda t-SNE zelo zmogljiva, je hkrati tudi dokaj počasna, zato si pomagamo s Barens-Hut aproksimacijo. Prav tako pa dimenzijo na začetku zmanjšamo s pomočjo PCA, šele nato pa uporabimo t-SNE.



Slika 3: Metoda t-SNE v dve dimezije.

Metoda t-SNE naredi projekcijo v dve dimenziji, kjer loči spektre, ki so drugačni od ostalih v posebne skupine. Opazimo skupine zvezd s Hα emisjo, skupine trojnih zvezd, in še nekaj izoliranih skupin, ki pa nevem katerim skupinam pripadajo.

Zanima nas kako vpliva izključitev Hα območja iz spektra na razporeditev emijskih zvezd. Iz spektra izvzamemo prvih 500 točk in spet naredimo t-SNE projekcijo (slika 4).



Slika 4: Metoda t-SNE v dve dimezije brez prbih 500 točk spektra.

Emijske zvezde so še vedno v posbnih gručah, kar pomeni da je tudi ostali del spektra emijsikih zvezd dovolj različen od ostalih, da jih metoda s-TNE izolira.

4 Zaključek

Pri tej nalogi smo se spoznali z dvema metodama za redukcijo dimenzij PCA in t-SNE. V tej nalogi se je metoda t-SNE izkazala za bolj praktično.