# Sloan Digital Sky Survey Classification

## Klasifikacija galaksija, zvezda i kvazara na osnovu DR14 iz SDSS-a

Student: Milica Petrović 2022/3267

### *O radu*

U ovom radu pokušaćemo da klasifikujemo opservacije svemira na galaksije, zvezde i kvazare. Prvo će biti opisana baza podataka koje želimo da klasifikujemo, zatim sledi analiza obeležja i kreiranje modela mašinskog učenja za predviđanje novih podataka.

Baza podataka koja se koristi u ovom radu je Sloan Digital Sky Survey (izdanje 14).

### *Baza podataka*

Sloan Digital Sky Survey je projekat koji pruža javne podatke o opservacijama svemira. Ova merenja se vrše od 1988 i dostupna su svima.

U tu svrhu napravljen je specijalni teleskop prečnika 2.5m u opservatoriji Apache Point u Novom Meksiku, SAD. Teleskop koristi kameru od 30 CCD-čipova sa svakim od 2048x2048 tačaka slike. Čipovi su raspoređeni u 5 redova sa po 6 čipova u svakom redu. Svaki red posmatra prostor kroz različite optičke filtere (u, g, r, i, z) na talasnim dužinama od približno 354, 476, 628, 769, 925 nm.

Teleskop pokriva oko jedne četvrtine zemaljskog neba - stoga se fokusira na severni deo neba.

Baza podataka se sastoji od 10.000 merenja svemira, pri čemu je svako merenje opisano sa 17 kolona obeležja i jednom kolonom koja predstavlja klasu i identifikuje set obeležja kao galaksiju, zvezdu ili kvazar.

### *Opis obeležja*

Baza je rezultat spoja dve vrste obeležja: 'PhotoObj' – obeležja koja sadrže fotometrijske podatke i 'SpecObj' – obeležja koja sadrže spektralne podatke.

<u>'PhotoObj'</u>

- objid = Identifikator objekta
- ra = J2000 Right Ascension (r-band)
- dec = J2000 deklinacija (eng. Declination) (r-band)

Right Ascension (skraćeno ra) i Declination (skraćeno dec) su astronomske koordinate koje određuju pravac tačke na nebeskoj sferi (tradicionalno naziv za nebesku sferu je nebo) u ekvatorijalnom koordinantnom sistemu.

- u, g, r, i, z predstavljaju odziv 5 opsega (filtera) teleskopa
- run = broj pokretanja (eng. run number)
- rerun = dodatni broj (eng. rerun number)
- camcol = broj kolone (eng. camera column)
- field = broj polja (eng. field number)

Run, rerun, camcol i field su karakteristike koje opisuju polje unutar slike koju je napravio SDSS. Polje je u osnovi deo cele slike koji odgovara 2048x1489 piksela. Polje se može identifikovati po:

- broj pokretanja, koji identifikuje specifično skeniranje,
- kolona kamere, ili 'camcol', broj od 1 do 6, koji identifikuje liniju skeniranja tokom akvizicije i
- broj polja. Broj polja obično počinje od 11 (nakon početnog vremena povećanja), a može biti i do 800 za posebno duga snimanja.
- Dodatni broj, rerun, određuje kako je slika obrađena.

'SpecObj'

- specobjid = Identifikator objekta
- class = klasa objekta (galaksija, zvezda ili kvazar)

Klasa identifikuje da je objekat galaksija, zvezda ili kvazar. Ovo će biti ciljna varijabla koju ćemo pokušati da predvidimo.

- redshift = crveni pomak
- plate = plate number
- mjd = MJD opservacije
- fiberid = fiber ID

U fizici, crveni pomak se dešava kada svetlost ili drugo elektromagnetno zračenje iz objekta poveća talasnu dužinu ili se pomeri ka crvenom delu spektra.

Svaka spektroskopska ekspozicija koristi veliku, tanku, kružnu metalnu ploču, pri čemu svaka ploča ima jedinstveni serijski broj – plate.

Modifikovani julijanski datum (eng. Modified Julian Date), koji se koristi za označavanje datuma snimanja određenog dela SDSS podataka (slika ili spektar).

SDSS spektrograf koristi optička vlakna za usmeravanje svetlosti prilikokm akvizicije u fokalnu ravan od pojedinačnih objekata do proreza. Svakom objektu se dodeljuje odgovarajući fiberID.

***Analiza obeležja***

Najpre ćemo pogledati osnovne karakteristike obeležja.

| | objid | ra | dec | u |
|---|---|---|---|---|
| Ukupan broj obeležja | 10 000 | 10 000 | 10 000 | 10 000 |
| Srednja vrednost | 1.237650e+18 | 175.529987 | 14.836148 | 18.619355 |
| Standardna devijacija | 0.000000e+00 | 47.783439 | 25.212207 | 0.828656 |

| | g | r | i | z |
|---|---|---|---|---|
| Ukupan broj obeležja | 10 000 | 10 000 | 10 000 | 10 000 |
| Srednja vrednost | 17.371931 | 16.840963 | 16.583579 | 16.422833 |
| Standardna devijacija | 0.945457 | 1.067764 | 1.141805 | 1.203188 |

| | run | rerun | camcol | field |
|---|---|---|---|---|
| Ukupan broj obeležja | 10 000 | 10 000 | 10 000 | 10 000 |
| Srednja vrednost | 981.034800 | 301.0 | 3.648700 | 302.380100 |
| Standardna devijacija | 273.305024 | 0.0 | 1.666183 | 162.577763 |

| | specobjid | redshift | plate | mjd | fiberid |
|---|---|---|---|---|---|
| Ukupan broj obeležja | 10 000 | 10 000 | 10 000 | 10 000 | 10 000 |
| Srednja vrednost | 1.645022e+18 | 0.143726 | 1460.986400 | 52943.533300 | 353.069400 |
| Standardna devijacija | 2.013998e+18 | 0.388774 | 1788.778371 | 1511.150651 | 206.298149 |

Iz datih tabela uočava se da nijedno obeležje nema nedostajućih vrednosti, što znači da nema dodavanja vrednosti obeležja. Baza podataka je potpuna.

Takođe primećuje se da većina karakteristika ostaje u razumnoj skali kada se uporede vrednosti unutar samo jedne kolone.

Primetno je i to da nema kategoričkih obeležja, osim klase, koja će kasnije biti kodirana odgovarajućim vrednostima.

Raspodela broja opservacija za svaku klasu prikazana je u sledećoj tabeli:

| Klasa | Broj opservacija |
|---|---|
| Galaksija | 4998 |
| Zvezda | 4152 |
| Kvazar | 850 |

Odavde se vidi da većina opservacija (oko 50%) pripada klasi galaksija, oko 40% klasi zvezda i malo manje od 10% pripada klasi kvazar, odakle zaključujemo da su klase nebalansirane.

Izdvajanje prediktora

U datom skupu prediktora pojedini prediktori nisu toliko povezani sa ciljnom promenljivom 'class' i iz tog razloga takva obeležja će biti izbačena iz originalnog skupa obeležja i neće se koristiti za klasifikaciju.

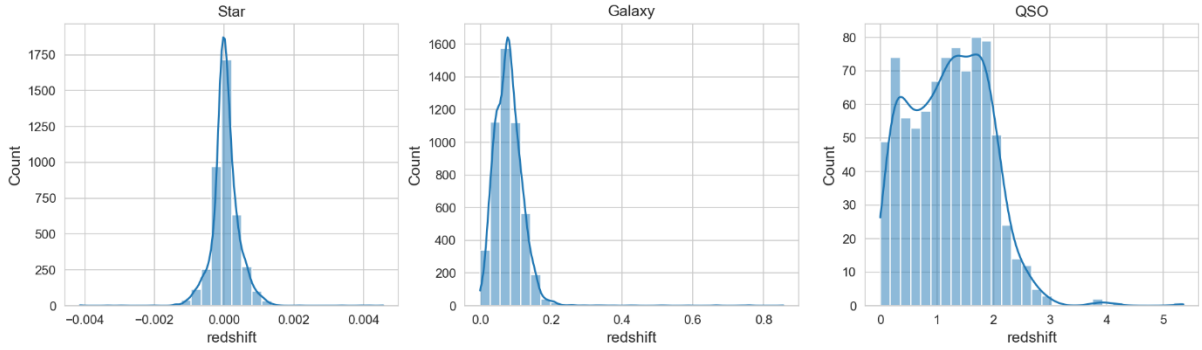
'objid' i 'specobjid' su samo identifikatori za pristup redovima kada su bili uskladišteni u originalnoj bazi podataka. Stoga nam neće biti potrebni za klasifikaciju jer nisu povezani sa ishodom.

Čak i više: karakteristike 'run', 'rerun', 'camcol' i 'field' su vrednosti koje opisuju delove kamere u trenutku kada se vrši posmatranje, npr. 'run' predstavlja odgovarajuće skeniranje koje je uhvatilo objekat.

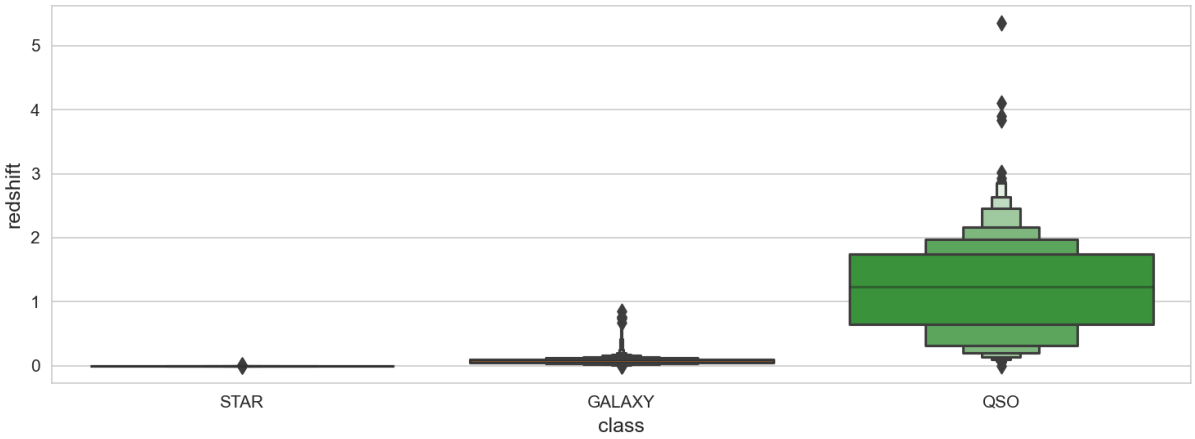
Navedena obeležja izbacujemo, kako bi svaka korelacija sa ciljnom varijablom bila slučajna.

'Redshift'

Da bismo analizirali prediktor 'redshift' prikazaćemo histogram vrednosti prediktora za svaku klasu posebno. Ovi histogrami će nam reći kako su vrednosti prediktora raspodeljene po klasama.




Takođe prikazaćemo i estimiranu raspodelu vrednosti prediktora po klasama.

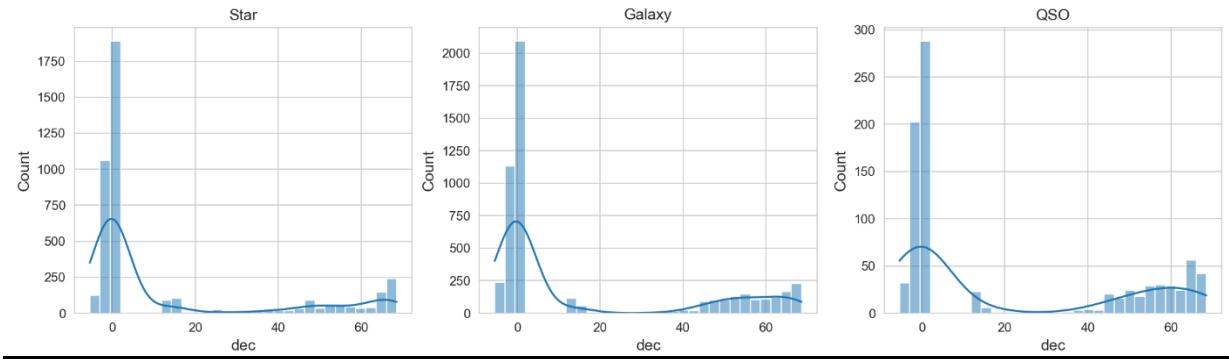



Sa prikazanih procena raspodele vrednosti obeležja prediktora uočavamo sledeće:

- Zvezda: Raspodela liči na normalnu raspodelu sa dosta malom standardnom devijacijom, i gotovo sve vrednosti su jednake nuli ili su veoma bliske nuli.
- Galaksija: Vrednosti prediktora 'redshift' mogu doći iz blago pomerene udesno normalne distribucije koja je centrirana oko 0.075.
- Kvazar: Vrednosti prediktora 'redshift' za kvazare su mnogo uniformnije raspoređene nego vrednosti kod zvezda ili galaksija. One su otprilike ravnomerno raspoređene od 0 do 3, a zatim se pojave drastično smanjuju. Za vrednosti bliske 4 i 5.5 pojavljuju se neki outlier-i.

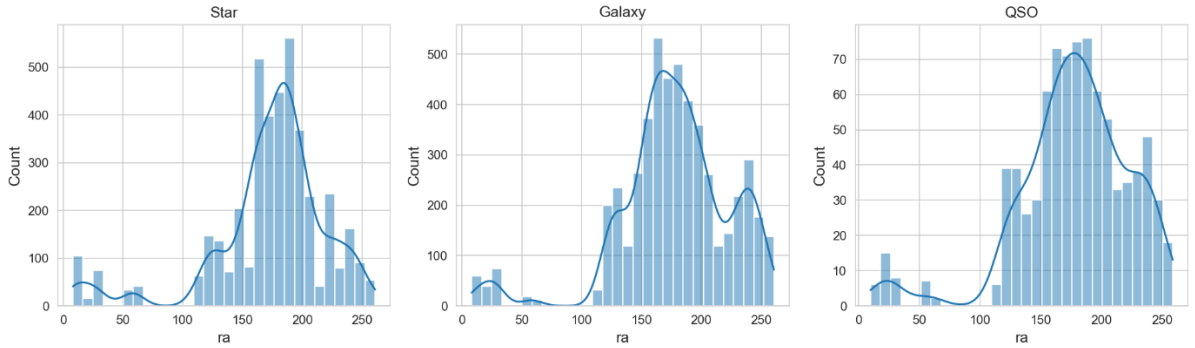
'dec'

Estimirana raspodela vrednosti prediktora 'dec' po klasama:



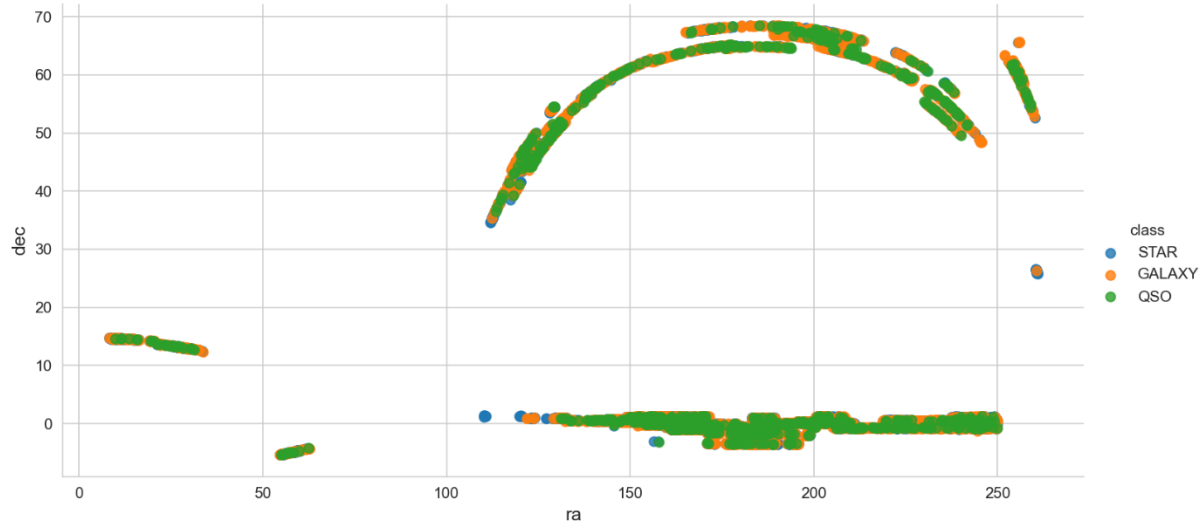

Sa datih histograma uočavamo da su vrednosti prediktora 'dec' gotovo identično raspodeljene u slučaju sve tri klase. Najviše ima vrednosti koje su bliske nuli ili nula, a pojavljuju se i manje-više ravnomerno raspodeljene vrednosti u opsegu od 45-70.

'ra'




Posmatranjem histograma vrednosti predikotra 'ra' takođe uočavamo, kao i kod prediktora 'dec', da opseg vrednosti u kojima se pojavljuje prediktor su prilično slični sa manjim odstupanjima.

Kako su 'ra' i 'dec' astronomske koordinate za određivanje položaja objekata u svemiru, biće prikazana zavisnost jedne od druge koordinate po klasama.

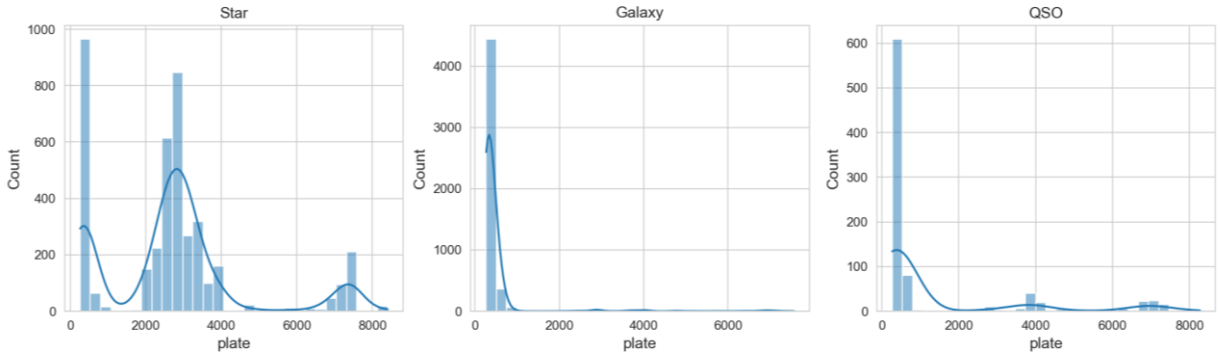



Kao što možemo jasno primetiti, astronomske koordinate se ne razlikuju značajno između 3 klase. Postoje neka odstupanja za zvezde i galaksije, ali za veći deo koordinate su unutar istog opsega.

Zašto je to?

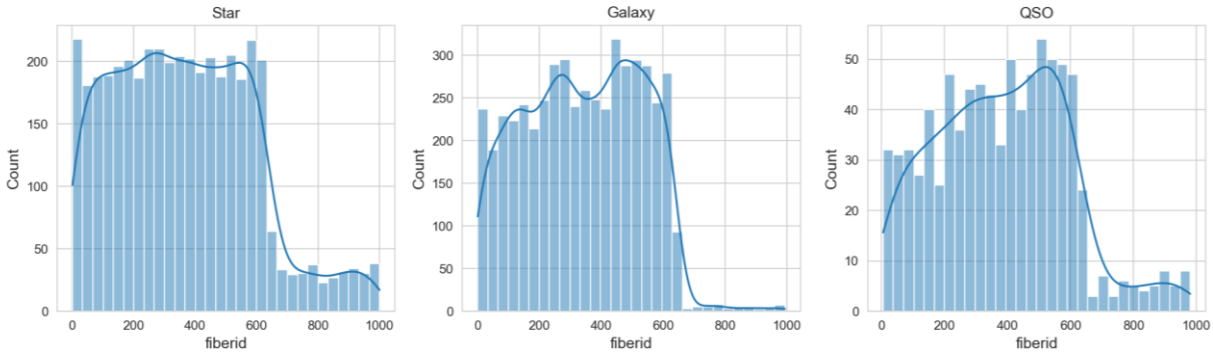
Sve SDSS slike pokrivaju istu oblast neba. Gornja slika nam govori da se zvezde, galaksije i kvazari posmatraju podjednako na svim koordinatama u ovoj oblasti. Dakle, gde god da SDSS „pogleda“ – šansa za posmatranje zvezde ili galaksije ili kvazara je uvek ista.

'plate'



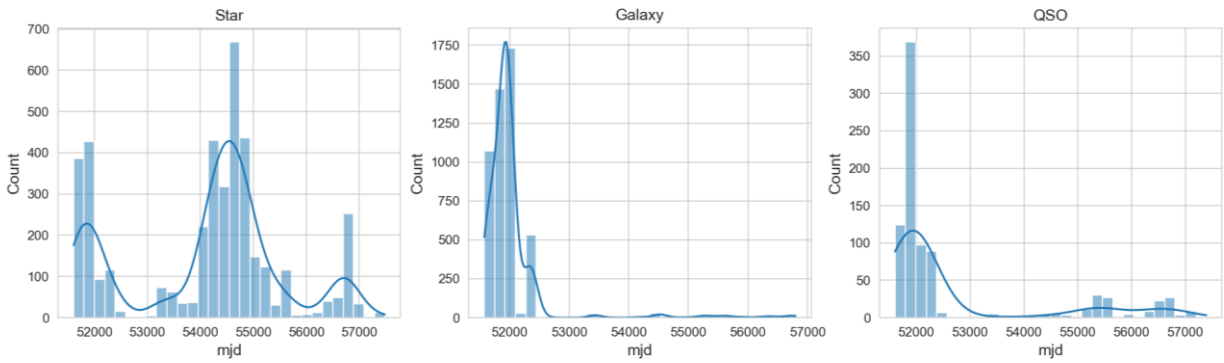

Posmatranjem histograma primećuje se da vrednosti prediktora za kalse galaksija i kvazar su poprilično slične. Skoncentrisane su oko vrednosti 500, sa manjim odstupanjima na većim vrednostima. Kod vrednosti prediktora klase zvezda primećuje se da pored toga što ima dosta vrednosti oko nule, kao i druge dve klase, ima značajan broj vrednosti i u opsegu 2000 – 4000. Pojavljuje se i manji broj vrednosti na oko 7000.

'fiberid'




Sve tri raspodele vrednosti prediktora su prilično uniformne u opsegu vrednosti od 0 do 600, nakon čega raspodela naglo opada na nulu u slučaju klase galaskija, do kod druge sve klase para ma manju konstantnu vrednost.

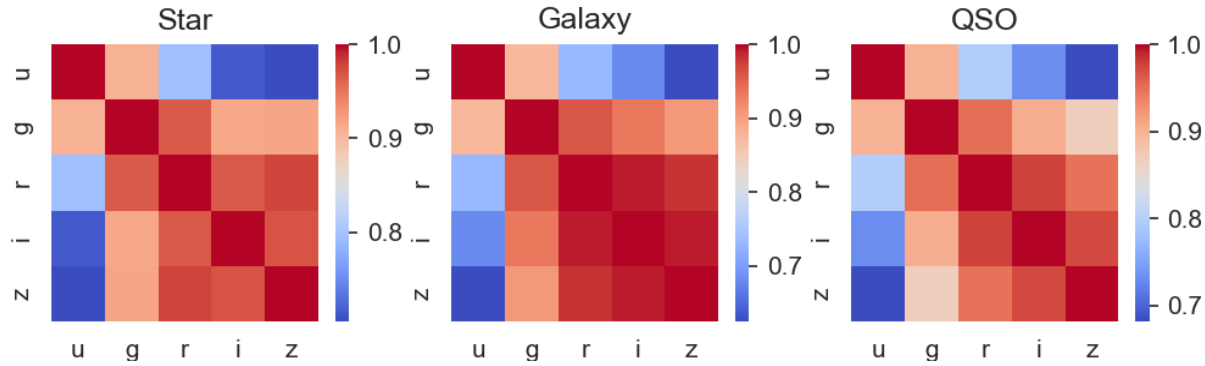
'mjd'



Histogrami klasa galaksija i zvezda su skoncentrisani oko vrednosti 52000, dok histogram klase zvezda, pored velikog broja primera sa vrednošću prediktora 52000, ima izraženu Gausovu raspodelu  sasrednjom vrednosti 54500. Takođe pojavljuju se i vrednosti bliske 57000.

'u', 'g', 'r', 'i', 'z'

'u', 'g', 'r', 'i', 'z' predstavljaju različite talasne dužine koje se koriste za snimanje opservacija. Posmatraćemo međusobnu korelaciju ovih prediktora.




Primećuje se da korelacione matrice izgledaju veoma slično za svaku klasu. Može se reći da postoje visoke korelacije između različitih opsega. Stoga je zanimljivo videti da je prediktor 'u' manje korelisan sa ostalim predikotrima. u, g, r, i, z hvataju svetlost na talasnim dužinama od oko 354, 476, 628, 769 i 925 nm. Ali, kao što vidimo - korelacija je otprilike ista za svaku klasu.

Redukcija dimenzija prediktora

Sada ćemo smanjiti broj dimenzija zamenom različitih opsega 'u', 'g', 'r', 'i' i 'z' linearnom kombinacijom sa samo 3 dimenzije koristeći analizu glavnih komponenti (PCA analiza).

PCA:

n opservacija sa p prediktora mogu se tumačiti kao n tačaka u p-dimenzionalnom prostoru. PCA ima za cilj da projektuje ovaj prostor u k-dimenzionalni podprostor (sa k<p) sa što manjim gubitkom informacija.

To radi pronalaženjem k pravaca u kojima n tačaka najviše variraju (-> glavne komponente). Zatim projektuje originalne tačke podataka u k-dimenzionalni podprostor. PCA vraća matricu nxk dimenzija.

Korišćenje PCA na našim podacima će smanjiti količinu operacija tokom obuke i testiranja.

$$PCA('\mathrm{u}','\mathrm{g}','\mathrm{r}','\mathrm{i}','\mathrm{z}') \rightarrow' PCA_1',' PCA_2',' PCA_3'$$

***Modeli mašinskog učenja***

Sada će biti trenirani različiti modeli na ovom skupu podataka.

Pre samog treniranja modela potrebno je izvršiti normalizaciju obeležja tako da se vrednosti svih obeležja nalaze u opsegu od 0 do 1.

Zatim je potrebno podeliti ceo skup podataka na trenirajući i testirajući skup, pri čemu je odnos ova dva skupa 80:20, respektivno. Modeli će biti obučavani pomoću trenirajućeg skupa podataka i testirani pomoću testirajućeg skupa podataka.

Naivni Bajes

Naivni Bajes spada u generativne algoritme. Pretpostavke koje Naivni Bajes pravi:

- Klase imaju Gausovu raspodelu
- Uslovnu nezavisnost prediktora pod uslovom da je poznata klasa

Ovaj klasifikator najpre određuje modele klasa, odnosno raspodelu $P(x|y)$ i apriornu verovatnoću pojave svake od klase $P(y)$. Zatim konačnu odluku donosi kao:

$$\hat{y} = arg\max_i P(y_i|x)P(y_i)$$

Tačnost klasifikacije koja se dobije korišćenjem ovog klasifikatora kros-validacijom je 97.7375%.

Metod nosećih vektora

Metod nosećih vektora klasifikuje podatke na osvnou separacione prave oblika $y = \langle w, x\rangle + b$, gde se vektor $w$ dobija pomoću nosećih vektora. Noseći vektori su oni primeri $x$ kod kojih je funkcionalana margina manja ili jednka 1.

Za parametar regularizacije $C = 350$ i polinomijalni kernel tačnost koja se ostvaruje na testirajućem skupu je 98.9625%.

Random Forest klasifikator

Slučajna šuma (eng. Random Forest) je algoritam koji se sastoji od više stabala, pri čemu svako stablo može a ne mora da se obučava nad istim skupom podataki, ali se sva stabla obučavaju nad podacima iz istog problema. Kao konačan rezultat klasifikacije slučajna šuma će dati usrednjen rezultat svih stabala.

Za broj stabala 50 tačnost koja se ostvaruje na testirajućem skupu kros-validacijom 99.1875%.

XGBoost klasifikator

XGBoost je zasnovan na Gradient Boost algoritmu sa nekoliko značajnih razlika koje ovaj algoritam čine izuzetno moćnim. Neke od prednosti koje XGBoost ima su:

- regularizacija, koja sprečava i preobučavanje

- paralelno procesiranje
- može da se izbori sa nedostajućim vrednostima

Tačnost koja se ostvaruje na obučavajućem skupu za 40 stabala je 99.35%.

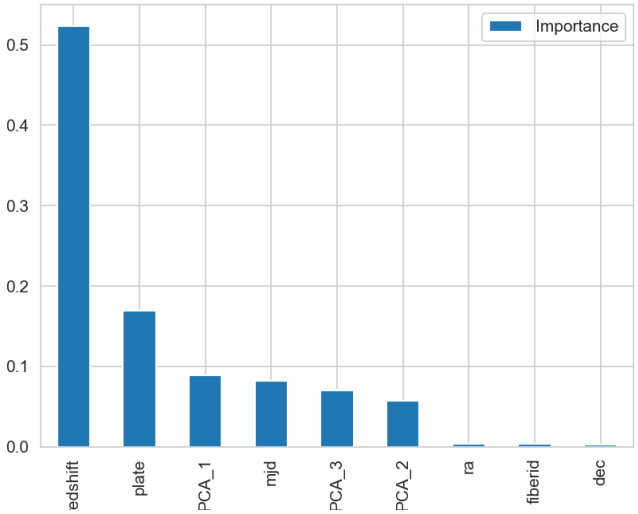
Tabelarni prikaz tačnosti klasifikatora na testirajućem skupu:

| Klasifikator | Tačnost |
|---|---|
| Naivni Bajes | 97.7375 % |
| Metod nosećih vektora | 98.9625 % |
| Slučajna šuma | 99.1875 % |
| XGBoost | 99.35 % |

Iz ove tabele primećuje se da su kod svih klasifikatora skoro svi primeri ispravno klasifikovani, gde najlošiju tačnost ima Naivni Bajesov klasifikator, a najbolju tačnost ima XGBoost.

Slučajna šuma ima lepu osobinu da rangira prediktore prema značaju za klasifikaciju, pa je prikazan njihov značaj sldećom tabelom i grafikom:

| Obeležje | Značaj |
|---|---|
| redshift | 0.523582 |
| plate | 0.168906 |
| PCA_1 | 0.088709 |
| mjd | 0.081678 |
| PCA_3 | 0.069768 |
| PCA_2 | 0.056933 |
| ra | 0.003759 |
| fiberid | 0.003536 |
| dec | 0.003129 |



Iz priložene tabele i grafika primećuje se da najveći uticaj na klasifikaciju imaju 'redshift' i 'plate'. Najmanji značaj imaju 'ra', 'fiberid' i 'dec', što odgovara gore prikazanim histogramima. Vrednosti obeležja po klasama imaju prilično istu procenu raspodele, pa je dobijeni rezultat i očekivan. Sada ćemo izbaciti ova tri obeležja i ponovo obučiti pomenute klasifikatore. Rezultati koji se tada dobiju koristeći isprojektovane klasifikatore su sledeći:

| Klasifikator | Tačnost |
|---|---|
| Naivni Bajes | 97.7625 % |
| Metod nosećih vektora | 98.96245 % |
| Slučajna šuma | 99.2125 % |
| XGBoost | 99.375 % |

Tačnost klasifikacije je ista bez ova tri obeležja kao i sa njima. Zaključak je da ih treba izbaciti i na taj način smanjujemo dimenzionalnost prostora obeležja, a samim tim i brži rad algoritma.

Konfuziona matrica za XGBoost klasifikator nad ovako izabranim prediktorima ima sledeći oblik:

| | $\widehat{galaksija}$ | $\widehat{zvezda}$ | $\widehat{kvazar}$ |
|---|---|---|---|
| $galaksija$ | 4964 | 26 | 8 |
| $zvezda$ | 29 | 820 | 1 |
| $kvazar$ | 6 | 0 | 4146 |

***Zaključak***

Baza podataka je veoma bogata i prediktori su dobri deskriptori klasa. Na osnovu prediktora mogu se sa visokom tačnošću (preko 99%) klasifikovati objekti.

Pokušati smanjivati broj obeležja prema značaju sve dok se tačnost klasifikacije ne pogoršava.