Seminarski rad iz predmeta Istraživanje podataka 1
Podaci: FIFA19

Nikola Janković
e-mail: nikola_jankovic@tuta.io

Matematički fakultet, Univerzitet u Beogradu

15. avgust 2019

Sažetak

Tema ovog seminarskog rada je detaljnija analiza, sa klasterovanjem, skupa podataka dobijenog iz poslednje verzije igrice FIFA19 (u toku pisanja rada) preuzetog sa adrese: `https://www.kaggle.com/karangadiya/fifa19`. U radu će biti prikazani neki od algoritama za klasterovanje i rezultati dobijeni primenom na ovaj skup. Uz pokušaj da budu dobijeni klasteri to približniji nekim podelama koje postoje u fudbalu.

1 Uvod

Klaster analiza je grupisanje objekata koje se oslanja samo na informacije koje se nalaze u podacima koji opisuju te objekte i veze među njima. Cilj klaster analize je da objekti u grupi budu slični(povezani) međusobno i drugačiji od objekata u drugim grupama. Što je veća sličnost u grupi i različitost među grupama klaster analiza je izrazitija.
U ovom seminarskom radu biće prikazani rezultati klaster analize pomoću algoritama koji su viđeni na kursu Istraživanje podataka 1:

- K-means
- DBSCAN
- Self Organizing Maps (*Kohonen*)
- Hijerahijsko klasterovanje

kao i dva dodatna algoritma:

- Mean-Shift
- BIRCH

Svi algoritmi su primenjeni uz pomoć biblioteka jezika Python uz korišenje softvera IBM Spss Modeler zbog loše dokumentacije vezane za SOM[1] u modulu *minisom*[2].

1.1 Skup podataka korišćen u radu

Skup podataka sastoji se od $\approx$ 18000 igrača(*slogova*) i 89 ocena(*atributa*). Blagi uvid u tabelu je moguć na slici 1.

ID	Name	Age	Photo	Nationality	Flag	Overall	Po
158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94
20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94
190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93
193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93
192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92

Slika 1: data.csv

[1] Self Organizing Maps
[2] `https://github.com/JustGlowing/minisom`

Podaci iz skupa se koriste kao parametri koje koristi kompanija *EA Sports* pri kreiranju simulacije fudbalera iz realnog sveta kako bi napravili distinkciju među njima.
Nazivi kolona uglavnom nedvosmisleno ukazuju na njihovo značenje, ali će ipak biti data objašnjenja za neke od atributa, koje korisnik smatra da nisu poznati većini.

- *Value* - Predstavlja procenjenu trenutnu vrednost igrača u dolarima, potrebno je praviti razliku u odnosu na atribut *Release Clause*
- *International Reputation* - Broj između 0 i 1 koji govori koliko je uspeha imao u igrama za reprezentaciju svoje zemlje.
- *Loaned from* - Pojedini igrači mogu biti posuđeni timu X od strane tima Y. Do kraja posudbe tim X je u obavezi da plaća igrača u istom iznosu kao što je to radio tim Y. Na kraju posudbe tim X ima prednost (u nekim slučajevima i pravo) da otkupi u potpunosti prava na igrača od tima Y.
- *LS, ST, RS, ..., RB* - Atributi koji predstavljaju koliko je projektovana Overall ocena igrača u slučaju da ga osoba koja igra igricu postavi na poziciju sa tim nazivom kolone.
- *Release Clause* - Procenjena cena koju je potrebno da tim Y plati timu X kako bi otkupio prava na igrača, često je vrednost ovog atributa veća u odnosu na atribut *Value*, pogotovo kod mlađih igrača.

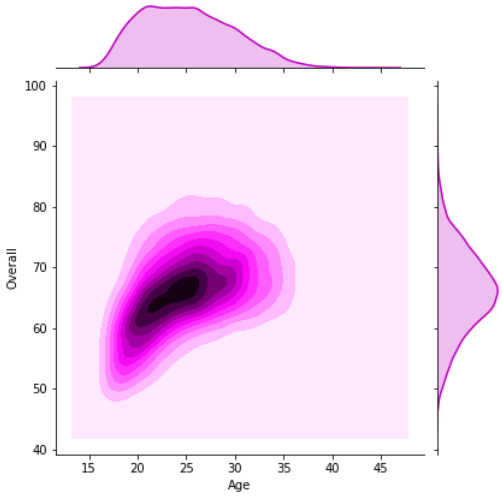
2 Analiza podataka

U ovom delu biće izložene neke zanimljive statistike iz skupa i prikazano kako je izvršeno pretprocesiranje podataka

2.1 Statistike

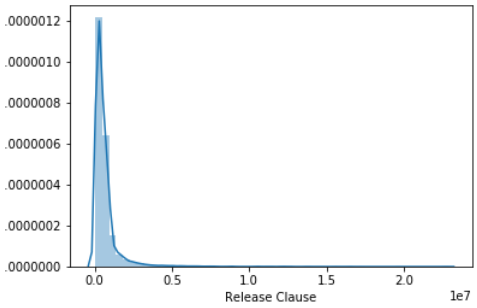
Na slici 2. možemo videti dijagaram zajedničke gustine raspodele za ocene *Overall* i *Age*

Vidimo da je raspodela za *Overall* normalna, dok *Age* podseća na neku $\tilde{\chi}^2$ raspodelu. Kao i da najveći broj fudablera ima između 23 i 27 godina sa *Overall* od 60 do 70 (potpuno očekivano).



Slika 2: Dijagram odnosa Age $\sim$ Overall

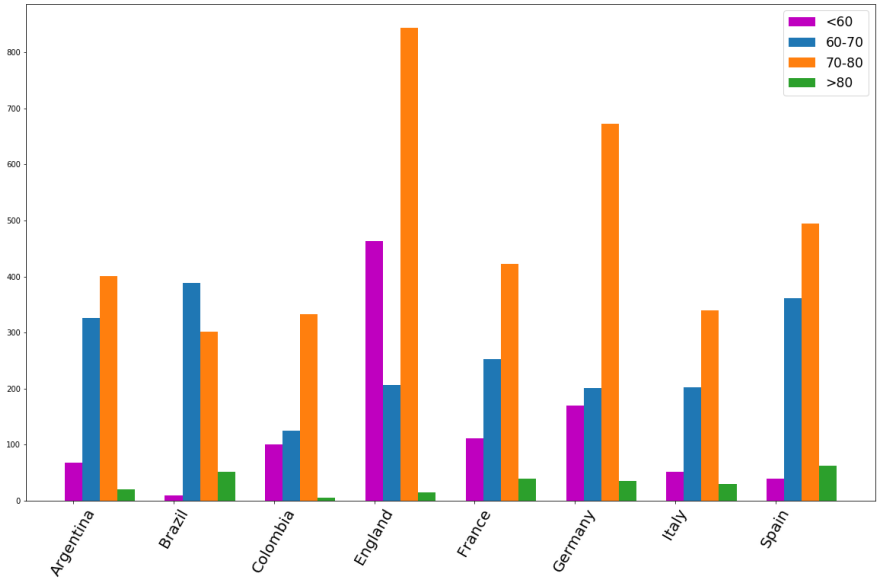
Druga zanimljiva statistika nam pokazuje raspodelu za atribut *Release Clause* i lako se uočava da najveća količina novca figurira u malom procentu igrača. Dok je kod igrača koji nisu vrhunske klase to značajno manje.



Slika 3: Raspodela izlazne klauze

I treći dijagram nam pokazuje koliko igrača nam dolazi iz koje države (razmatrane samo države koje imaju više od 500 predstavnika)

Primetno je da je broj igrača iz Engleske najveći kao i da dominiraju u broju igrača sa ocenom manjom od 60. Razlog ovome je to što u igrici postoje timovi iz čak 4 engleska ligaška takmičenja u kojima većinu čine igrači iz Engleske, a kako su timovi iz 3. i 4. ranga polu-profesionalni, ocene za igrače su očekivano niske.



Slika 4:

2.2 Pretprocesiranje