Napoved NBA tekem

Seminarska naloga pri predmetu Umetna inteligenca

Mentor: as. mag. Petar Vračar

Avtorja: Žan Bizjak(63170055)

Žiga Šolar(63170282)

Ljubljana, december 2018

Kazalo vsebine

[Uvod 3](#_Toc531545827)

[Priprava podatkov 4](#_Toc531545828)

[Klasifikacija 5](#_Toc531545829)

[Klasifikacijska točnost in brierjeva mera 5](#_Toc531545830)

[Naključni gozdovi 5](#_Toc531545831)

[K najbljižjih sosedov 5](#_Toc531545832)

[Odločitveno drevo 5](#_Toc531545833)

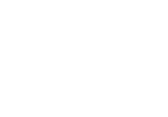
[Boosting 5](#_Toc531545834)

[Najivni Bayes 5](#_Toc531545835)

[Regresija 6](#_Toc531545836)

# Kazalo slik

[Slika 1: grafičen prikaz računanja povprečnih podatkov 4](#_Toc531545953)



# Uvod

Seminarska naloga govori o raziskovalni nalogi, kjer je bil cilj sestaviti klasifikacijske in regresijske modele, ki napovedujejo možnost zmage domače ekipe in razlika v točkah med dvema ekipama. Podani so nam bili podatki iz treh sezon košarkaške lige NBA. Te podatke je bilo potrebno primerno preurediti in na podlagi novih podatko zgraditi razne modele.

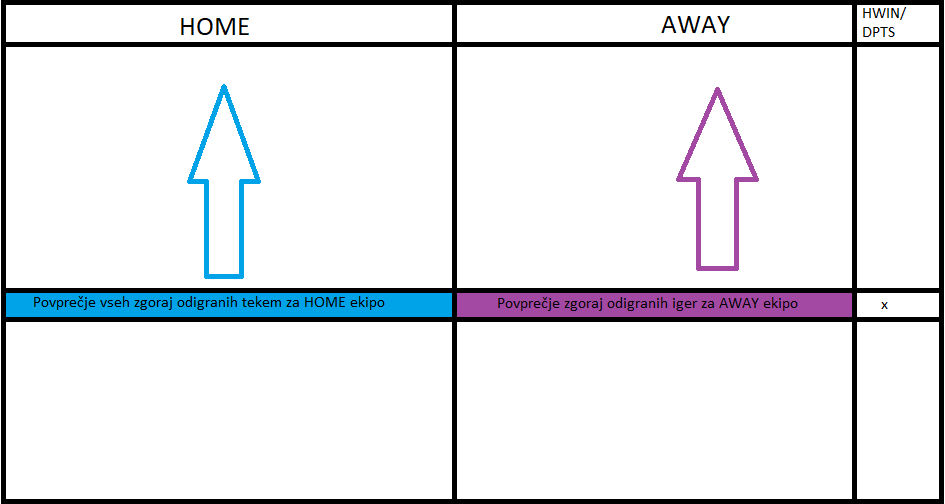
# Priprava podatkov

Cilj naše naloge je bil napovedati točke/zmagovalca. Za tako vrsto napovedi so podani podatki, taki kot so, neuporabni, saj bi se model naučil napovedati zmago za posamezno tekmo, in ne nasploh. Za to so se podatki prvo predelali v uporabne. Prvo, se je odstranil atribut SEASON, ker nam za učenje ne pomaga. Vse atribute, ki so bili podani kot »x akcij uspešnih, y akcij poskušenih«, so bili spremenjeni v skupni atribut, ki je ponazarjal uspešnost (npr. uspešnost meta za 3 pike). Iz podatka HPTS in APTS se je ustvaril nov atribut:

* (klasifikacija)HWIN: Ta ima vrednost 1, če je zmagala domača ekipa, in 0, če ne.
* (regresija)DPTS: Razlika točk med domačim in gostovalnim moštvom

Atributa HPTS in APTS, sta se po tem zbrisala, saj smo ju že porabili za določitev novega atributa.

Po preurejanju podatkov so spremenjeni v »uporabne« podatke, katere uporabimo za učenje in testiranje. Za vsako tekmo, se je izračunalo povprečje atributev vseh prejšnjih tekem, za določeno ekipo. Povprečja za prve tekme ne moremo zračunati, ker nimamo nobenih predhodnih podatkov. Zato se nova množica podatkov začne z 48. tekmo iz regular.txt podatkov. Po 48 tekmi v množici podatkov so vse ekipe vsaj enkrat igrale doma in vsaj enkrat igrale v gosteh. S temi podatki se nato naprej lotimo učenja modelov. Podatki so primerni, ker v primerjavi z prejšnjimi »surovimi« podatki sedaj ne gledamo podatke posamične tekme, ampak povprečja rezultatov do takrat odigranih tekem



Slika 1: grafičen prikaz računanja povprečnih podatkov

Iz poračunanih podatkov se množica razdeli na učno in testno množico. Učna zavzema prvih 2415 tekem, testna množica pa ostalo

# Klasifikacija

Pri klasifikaciji je število možnih opredelitev diskretno, zato se za razred uporabi HWIN. Za izgradnjo modelov so se je uporabilo 5 različnih algoritmov:

* Regresijsko drevo
* Najivni Bayes
* K najbljižjih sosedov
* Naključni gozdovi
* Boosting

## Klasifikacijska točnost in brierjeva mera

### Naključni gozdovi

Za ta primer se je izkazalo da je najbolj točen bil model zgrajen z RF algoritmom. In to z ~63,0% klasifikacijsko točnostjo. Brier score tega modela je bil ~0,455

### K najbljižjih sosedov

K najbljižjih sosedov ima CA ~ 61,7%, brierjevo mero pa ~0,517

### Odločitveno drevo

Odločitveno drevo ima CA ~ 61,1%, brierjevo mero pa ~0,498

### Boosting

Boosting ima CA ~ 62,0%, brierjevo mero pa ~0,467

### Najivni Bayes

Najivni Bayes ima CA ~ 62,2%, brierjevo mero pa ~0,503

# Regresija