Pràctica 1: Regressió

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

Introducció

En aquesta pràctica analitzarem una base de dades sobre els partits que ha jugat Michael Jordan. Veurem quines distribucions tenen els atributs, la relació entre ells i intentarem predir els punts que farà en un partit.

Apartat C

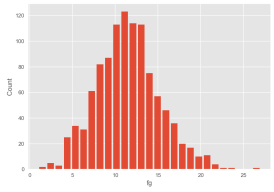
Els atributs que tenim a la base de dades són els següents:

0. Game [integer]: ID del partit.
1. Date [string]: data del partit en el format "yyyy/mm/dd".
2. Age [string]: edat del jugador mesurada en anys i dies.
3. Team [string]: equip al que pertany el jugador.
4. Opp [string]: equip adversari en el partit.
5. Result [string]: diferència de punts del equip respecte l'adversari, acompanyat d'una lletra Win/Lose.
6. mp [string]: minuts jugats amb el format "mm:ss"
7. fg [integer]: tirs anotats.
8. fga [integer]: tirs intentats.
9. fgp [float]: freqüència relativa de tirs anotats.
10. three [integer]: triples anotats.
11. threeatt [integer]: triples intentats.
12. threep [float]: freqüència relativa de triples anotats.
13. ft [integer]: tirs lliures anotats.
14. fta [integer]: tirs lliures intentats.
15. ftp [float]: freqüència relativa de tirs lliures anotats.
16. orb [integer]: rebots ofensius.
17. drb [integer]: rebots defensius.
18. trb [integer]: rebots totals.
19. ast [integer]: assistències.

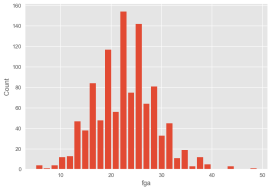
20. stl [integer]: pilotes robades
21. blk [integer]: llançaments de dos punts bloquejats.
22. tov [integer]: pilotes perdudes.
23. pts [integer]: punts aconseguits pel jugador.
24. game_score [float]: mitjana de punts de l'equip per jugador.
25. minus_plus [NaN]: diferència en el marcador respecte la entrada i la sortida del jugador.

No cal netejar massa la base de dades, només hi ha tres atributs on trobem *NaNs*. Els dos primers són *threep* i *ftp*, freqüències relatives de triples i tirs lliures respectivament, que no estan definits si no hi ha hagut cap intent (divisió per zero). L'últim atribut amb *NaNs* és *minus_plus*, que té tots els valors sense definir i, per tant, descartarem aquest atribut.

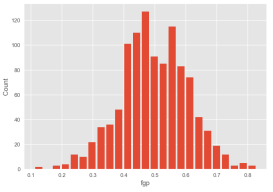
Per veure quins atributs segueixen una distribució Gaussiana hem fet histogrames i tests d'hipòtesi. A simple vista, els atributs que més s'assemblen a una distribució normal són els tres atributs pels tirs, els punts i el game_score. Hem fet els testos de normalitat a totes les variables però a les taules només veurem aquests cinc atributs.



Anotats

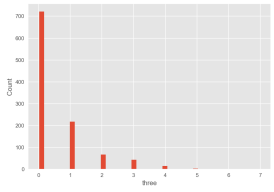


Intentats

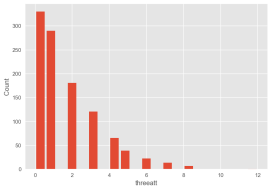


Freqüència relativa

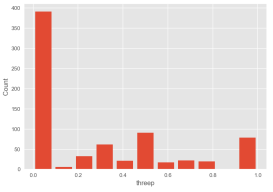
Figura 1: Tirs



Anotats

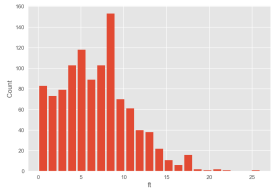


Intentats

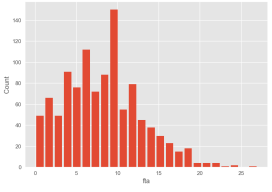


Freqüència relativa

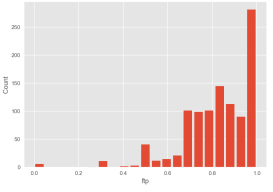
Figura 2: Triples



Anotats



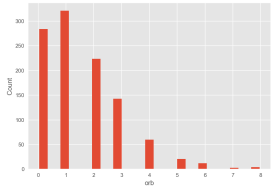
Intentats



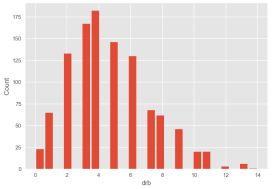
Freqüència relativa

Figura 3: Tirs lliures

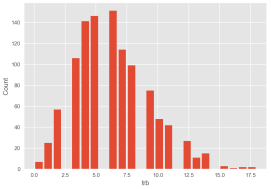
Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740



Ofensius

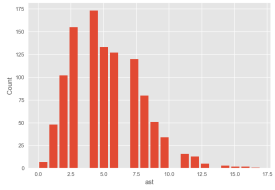


Defensius

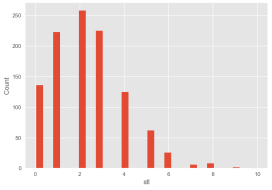


Totals

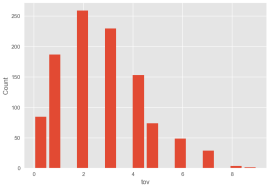
Figura 4: Rebots



Asistències

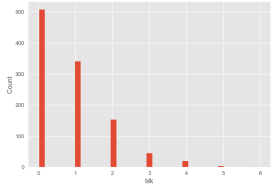


Boles robades

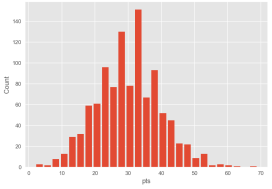


Boles perdudes

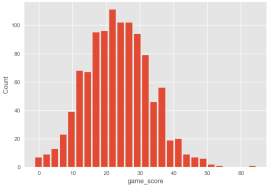
Figura 5: Asistències, boles robades i perdudes



Bloquejos



Punts



Game_score

Figura 6: Bloquejos, punts i game_score

Atribut	Estadístic Shapiro	p Phapiro	Estadístic Agostino K^2	p Agostino K^2
fg	0.989	0.000	14.586	0.001
fga	0.993	0.000	12.643	0.002
fgp	0.998	0.119	2.2689	0.261
pts	0.996	0.014	7.417	0.024
game_score	0.997	0.030	7.417	0.025

Taula 1: Estadístics de Shapiro i Agostino K^2 amb el p-valor corresponent

Per a que els tests d'hipòtesi corroborin que són distribucions gaussianes cal que $p > \alpha$, on α normalment val 0.05, per tant és fàcil veure que l'únic atribut amb distribució gaussiana és *fgp*. Tant *pts* com *game_score* es queden molt a prop de ser-ho.

Hem decidit que l'atribut objectiu serà els punts que farà Michael Jordan en cada partit ja que és un dels atributs més interesants si analitzem la carrera d'un jugador. Ha sigut difícil l'elecció ja que per predir aquest atribut amb una regressió cal tenir les dades dels demés, però, una vegada obtingudes aquestes dades, ja tindriem el valor que volíem predir. Es per això que la nostra predicció amb regressió serà només d'interès acadèmic i no tindrà gaire sortida pràctica.

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

Apartat B

Els atributs que són més importants per a una bona predicció són aquells que tinguin una correlació relativament alta. Com podem veure, aquests atributs són: *fg*, *fga*, *fgp*, *ft*, *fta* i *game_score*.

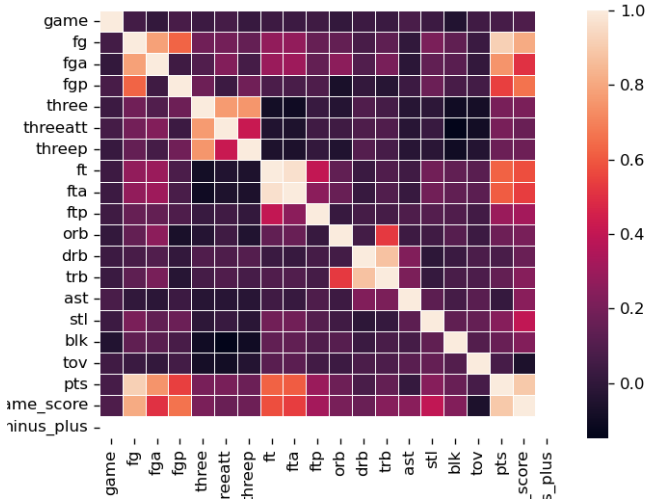


Figura 7: Heatmap de correlacions

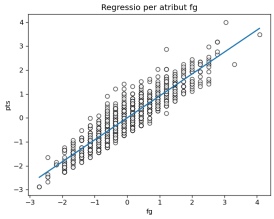
El fet de normalitzar les dades fa que els dominis dels atributs no desproporcionin la seva contribució al MSE, i això permet que funcioni millor el descens del gradient.

Hem fet un PCA i hem obtingut dues components principals. Per a fer-ho hem calculat la matriu de covariàncies pels atributs numèrics, i hem buscat els valors i vectors propis. Els valors propis ens diuen quanta informació hi ha en cada component: $w = (219.4, 29.6, 22.9, 15.3, 6.9, 4, 3.6, ...)$. Les dues primeres components tenen el 70,98% i el 9,56% d'informació respectivament.

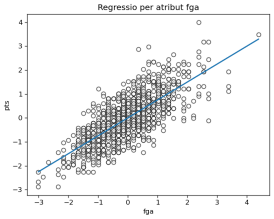
L'atribut que assoleix un MSE menor és el *fg*. Aquest resultat no sorpren a ningú ja que estem predint els punts que farà a partir dels tirs anotats. El millor MSE i R^2 score l'obtenim fent servir les dades transformades pel PCA. A la següent taula recollim el MSE i el R^2 score de les dades normalitzades si fem una regressió agafant només un atribut, agafant-los tots i agafant els dos primers atributs del canvi del PCA:

Atribut	MSE	R^2 score
fg	0.161090	0.849104
fga	0.435080	0.592454
fgp	0.787710	0.262141
ft	0.627369	0.412334
fta	0.639979	0.400523
ftp	0.949395	0.110688
game_score	0.219541	0.794353
Tots	0.100316	0.895684
PCA $n = 2$	0.092544	0.920129

Finalment, veurem algunes regressions. Només hem pogut fer-les unidimensionals, les regressions amb tots els atributs i amb els que ha proporcionat el PCA no les hem pogut dibuixar.



Tirs anotats



Tirs intentats

Freqüència relativa tirs

Figura 8: Regressions amb un atribut

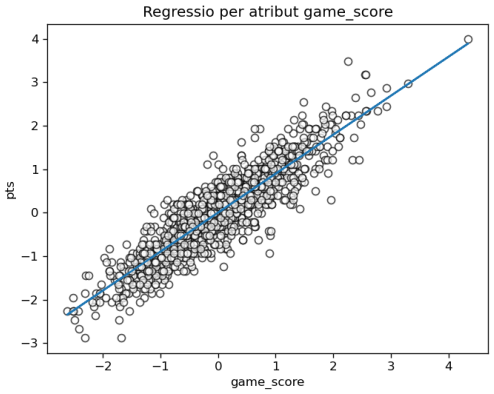


Figura 9: Regressió amb game_score