

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 [integer]: 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]`: **No entenem què vol dir aquest float.**
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 tests de normalitat a totes les variables però a les taules només veurem aquests cinc atributs.

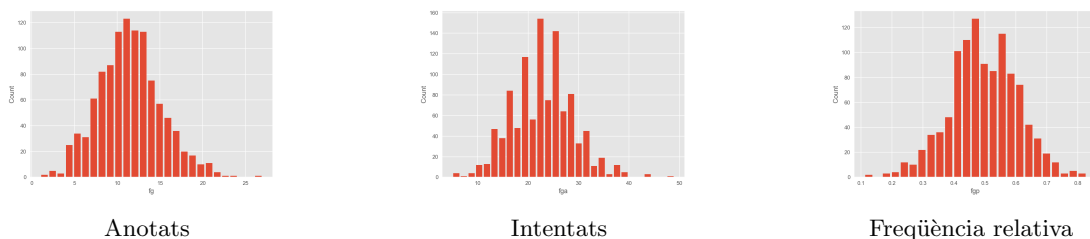


Figura 1: Tirs

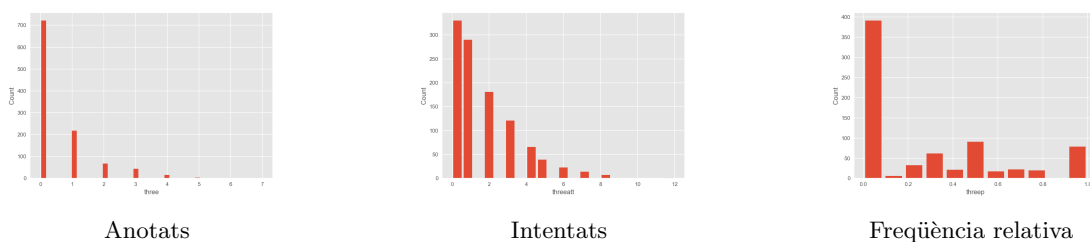


Figura 2: Triples

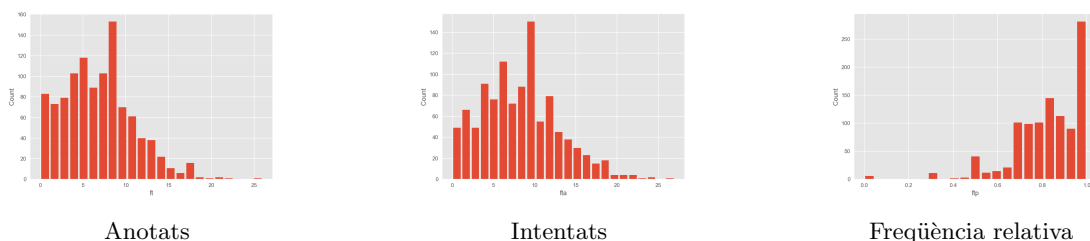


Figura 3: Tirs lliures

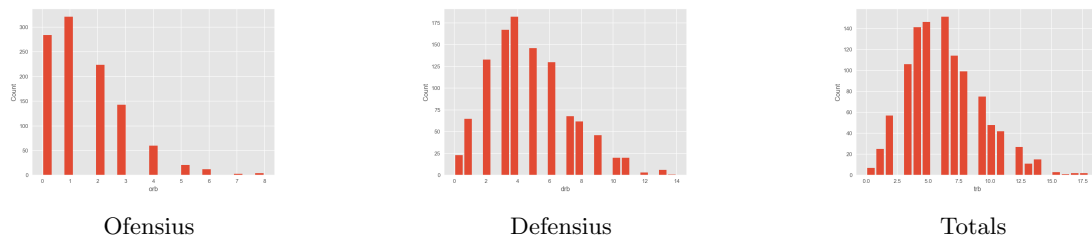


Figura 4: Rebots

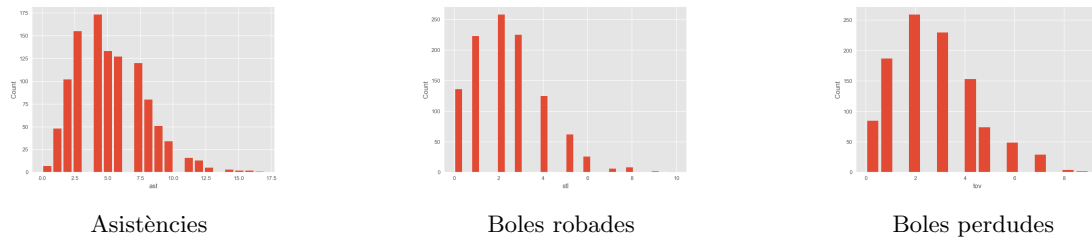


Figura 5: Assistències, boles robades i perdudes

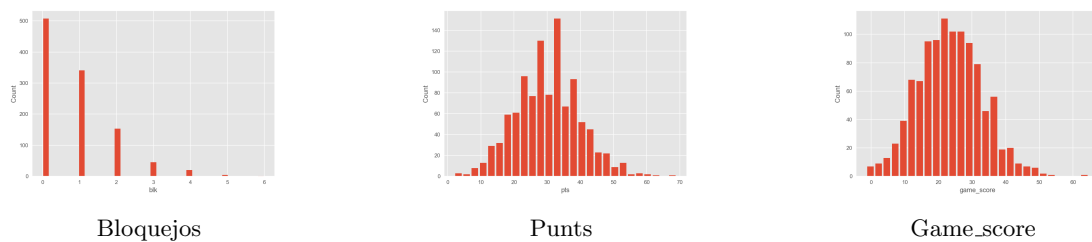


Figura 6: Bloquejos, punts i game_score

Atribut	Estadístic Shapiro	p Shapiro	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 interessants 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.

Apartat B