Universitat Autònoma de Barcelona
Facultat de Ciències

Qui és el millor?

Autora:

Ona Sánchez

1601181

16 de Desembre del 2022

Índex

1 Introducció

L'objectiu d'aquesta pràctica és, mitjançant la interfície proporcionada per Jupyter Notebook, estudiar i predir un valor en funció d'un conjunt de paràmetres que es calcularan mitjançant un conjunt de dades.

Les dades han sigut proporcionades per la web de Kaggle, concretament, la base de dades: *Michael Jordan, Kobe Bryant and Lebron James stats.*

L'objectiu d'aquesta base de dades és estudiar quin dels tres jugadors és el "millor". Així doncs, després d'un estudi de les dades s'ha decidit intentar predir el *"PER": Player Efficiency Rating* en funció de les altres variables, i veure quin dels 3 jugadors obté un valor més elevat.

El dataset que s'utilitza es pot trobar al següent enllaç:

https://www.kaggle.com/xvivancos/michael-jordan-kobe-bryant-and-lebron-james-stats/notebooks.

2 Presentació de les funcions

2.1 Llibreries i importacions

Per tal de poder dur a terme la nostra tasca és imprescindible tenir intal·lades les següents llibreries, ja que s'utilitzen les funcions següents (d'entre altres).

Llibreria	**Funció utilitzada**
sklearn.datasets	make_regression
numpy (as np)	shuffle isnan min max floor reshape array
pandas (as pd)	read_csv DataFrame
matplotlib pyplot (as plt)	figure subplots plot hist
seaborn (as sns)	heatmap pairplot
sklearn.linear_model	LinearRegression Lasso ElasticNet BayesianRidge
math	*operacions aritmètiques vàries*
sklearn.metrics	r2_score
ipywidgets	interact
mpl_toolkits.mplot3d	axes3d
intertools	combinations
plotly.express (as px)	scatter_matrix
sklearn.decomposition	PCA
sklearn.model_selection	RandomizedSearchCV train_test_split
scipy.stats	uniform

2.2 Funcions programades

2.2.1 standarize

- Entrada:
  - **np.array** X
- Sortida: **np.array** x
- Funcionament: Per cada atribut, calcula la mitjana i la desviació estàndar, posteriorment normalitza cada dada restant la mitjana i dividint per la desviació estàndar.
- Informació rellevant: Funció utilitzada exclusivament com a pas intermedi per a estudiar el dataset i millorar la seva predicció.

2.2.2 mse

- Entrada:
  - **np.array** y1
  - **np.array** y2
  - **int** inici
- Sortida: **float** mse
- Funcionament: Es comprova que la mida de y1 i y2 sigui igual i es calcula

$$\frac{1}{n}\sum_{i=1}^{n}(y_{1i}-y_{2i})^2$$

- Informació rellevant: La funció és utilitzada per tots els regressors per estudiar l'error que cometen.

2.2.3 regression

- Entrada:
  - **np.array** x
  - **np.array** y
- Sortida: Regressor lineal (*LinearRegressor()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor lineal simple com per al regressor multilineal simple.

2.2.4 split_data

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** train_ratio
- Sortida: **np.array** x_train, **np.array** y_train, **np.array** x_val i **np.array** y_val.
- Funcionament: mitjançant els mètodes *shuffle* i *floor* de la llibreria numpy creem les 4 llistes que retornem amb els components de les x i y aleatòriament ordenats i dividits.
- Informació rellevant: Funció utilitzada tant per validar el comportament dels regressors lineals com els multilineals i evitar el overfitting. En cas de no introduir el train_ratio, aquest tindrà el valor per defecte 0.8.

2.2.5 lasso

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** a
- Sortida: Regressor lineal (*Lasso()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió *Lasso* amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor multilineal amb el mètode de la llibreria sklearn *Lasso*. En cas de no introduir la a, aquesta tindrà el valor per defecte 0.1.

2.2.6 Bayes

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** t
- Sortida: Regressor lineal (*BayessianRidge()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió BayessianBrigde amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor multilineal amb el mètode de la llibreria sklearn *BayessianBrigde*. En cas de no introduir la t, aquesta tindrà el valor per defecte 10^{-6}.

2.2.7 ElasNet

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** t
- Sortida: Regressor lineal (*ElasticNet()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió ElasticNet amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor multilineal amb el mètode de la llibreria sklearn *ElasticNet()*. En cas de no introduir la t, aquesta tindrà el valor per defecte 10^{-6}.

3 Gestió del dataset

3.1 Explicació del Dataset

La base de dades es divideix en 7 datasets, dels quals s'ha usat i s'explicarà el dataset *advanced_ stats.csv*, ja que és el que conté la variable objectiu que es preten predir, *PER*.

El dataset conté les dades de diversos partits dels jugadors Michael Jordan, Lebron James i Kobe Bryant, sumant entre els 3 jugadors 92 partits i 29 atributs.

Per tant, el nostre dataset és de mida 92x29 (files x columnes).

Els 29 atributs recollits dels partits són els següents:

Atribut	**Explicació**	**Tipus de dada**
Season	Temporada	string
Age	Edad jugador	naturals [18-39]
Tm	Equip	string
Lg	Liga	string
Position	Posició	string
G	Jocs	naturals [3-82]
MP	Minuts jugats	naturals [128-3401]
PER	Rating d'eficiència	float [10.7-37.4]
TS%	Percentatge de tirs	float [0.47-0.67]
3PAr	Taxa d'intents d'anotar punts	float [0.01-0.42]
FTr	Intents de tir lliure	float [0.21-0.74]
ORB%	Percentatge de *rebound* ofensiu	float [0.5-6.8]
DRB%	Percentatge de *rebound* defensiu	float [7.8-24.3]
TRB%	Percentatge total de *rebound*	float [5.1-14.7]
AST%	Percentatge d'assistència	float [12.1-46.4]
STL%	Percentatge de robades	float [0.7-3.9]
BLK%	Percentatge de bloquejos	float [0.1-3.4]
TOV%	Percentatge de rotació	float [6.7-29.2]
USG%	Percentatge d'ús	float [22.5-39.2]
OWS	Victories compartides ofensives	float [-0.9-15.2]
DWS	Victories compartides defensives	float [-0.1-6.5]
WS	Victòries compartides	float [-0.4-21.2]
WS/48	Victòries compartides cada 48 minuts	float [-97-0.4]
OBPM	*Box* ofensiva	float [-4.7-14.8]
DBPM	*Box* defensiva	float [-3.3-5.8]
BPM	*Box Plus/Minus*	float [-5.9-18.2]
VORP	Valor per sobre del suplent	float [-0.2-12]
Player	Nom jugador	string
RSorPO	Tipus de temporada	string

3.2 Tractament de les dades

Estudiant les dades del dataset, s'observa que no hi ha incongruències entre els atributs, i que no hi ha cap tipus de dada faltant, pel que no cal fer tractament de nulls.

Es decideix, per tant, avançar al tractament de tipus de dades per tal de continuar l'estudi en la búsqueda del millor jugador.

Per la correcta gestió de les dades és necessari que totes tinguin valors numèrics (*int64* i *float64*), pel que caldria passar tots els valors de tipus *object* a *int64* i *float64*. Tot i això, abans de fer la transformació es decideix estudiar més a fons l'atribut objectiu *"PER"*, i es troba la següent fòrmula:

$$PER = PTS + 0.4 \cdot FG - 0.7 \cdot FGA - 0.4 \cdot (FTA - FT) + 0.7 \cdot ORB + 0.3 \cdot DRB + STL + 0.7 \cdot AST + 0.7 \cdot BLK - 0.4 \cdot PF - TOV \quad (1)$$

D'aquesta fòrmula [1] es dedueix que no és necessari cap dels atributs del tipus *string* per tal de predir el *Player Efficiency Rating*, pel que aquestes dades no calen ser modificades, ja que no es treballarà amb elles, i es decideix suprimir-les del *dataset*.

[1] Destacar que a la fòrmula tots els atributs que s'usen són numèrics i es troben en gran part al dataset *advanced_ stats.csv*

4 Estudi del dataset

4.1 Decisions preses abans d'estudiar el dataset

Per tal d'estudiar el dataset, es va decidir comparar els valors obtinguts per a la regressió lineal simple estandaritzant les dades i sense fer-ho. Una vegada vist si hi existeix millora o no amb l'estandaritzat escollim amb quina de les dues formes d'analitzar-lo ens quedem per a fer la resta d'anàlisis.

Per evitar treballar amb moltes dades, s'ha intentat crivar les variables per només usar les més significatives i, així, asseguar que utilitzem atributs que tenen una correlació directa (o prou alta) amb l'atribut objectiu (*PER*).

Per tal de veure si existeixen relacions entre les dades (així com si les dades segueixen alguna distriució), part de l'anàlisi s'ha centrat a observar histogrames de les dades i gràfics de relació dels atributs amb l'atribut objectiu. D'aquesta manera podem crivar dades que tenen distribucions que sabem que no ajuden a predir ni millorrar la predicció (com atributs que segueixen distribucions uniformes, etc...).

Per altra banda, s'ha plantejat el següent valor de criva per tal de discutir si un atribut és rellevant o no mitjançant la correlació amb l'atribut objectiu *PER*:

- Si aquest és superior a 0.5 diem que és rellevant per a predir el *PER*.

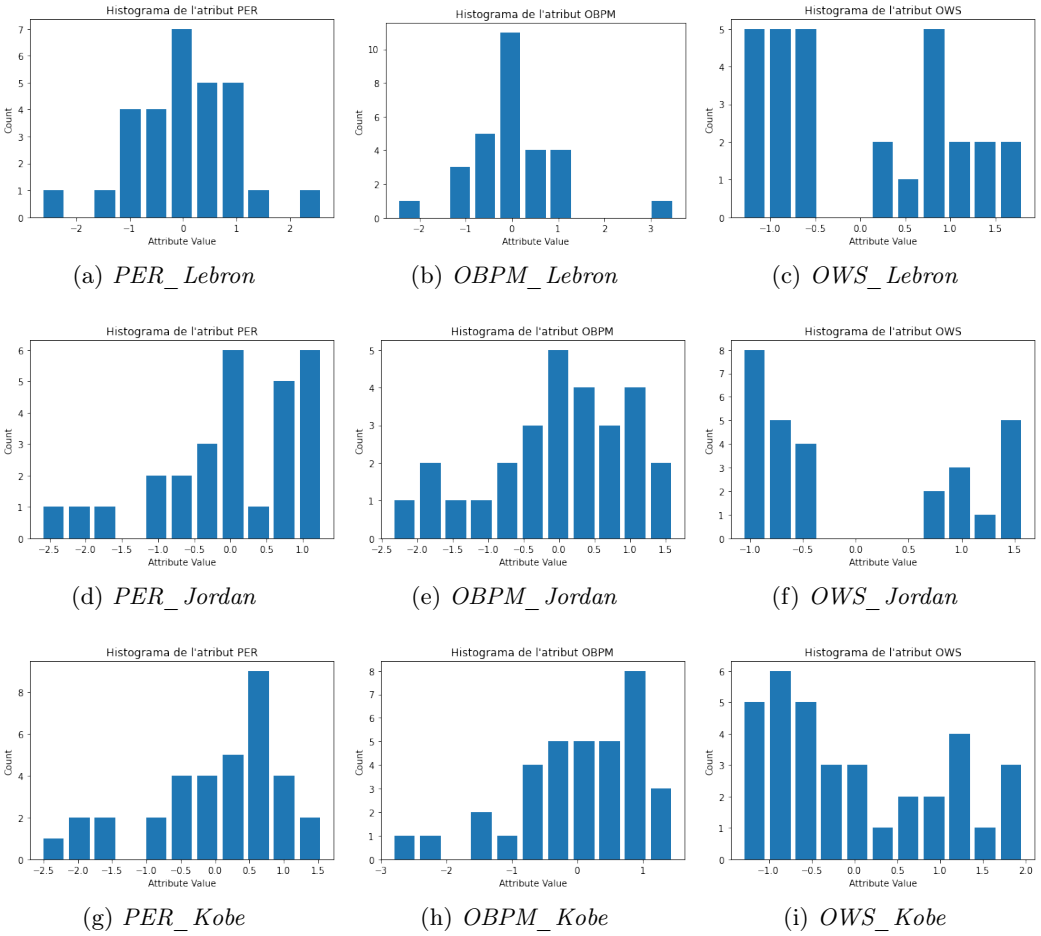
Fem servir aquest valor per assegurar que es treballa amb:

1. Valors significatius per a predir/millorar la predicció.
2. Valors fortament correlacionats amb el conjunt de les dades.

Cal destacar que, per tal de poder estudiar quin dels 3 jugadors és millor, s'ha decidit dividir el dataset en 3 sub-datasets: dataset_Lebron, dataset_Jordan i dataset_Kobe, pel que, a partir d'aquest punt, cada part de l'estudi de la predicció (estudi de les distribucions de les dades, millors regressors, etc.) s'ha fet per cadascun dels jugadors, per tal de comparar els resultats (millor *PER*) i així poder concloure quin és el millor.

4.2 Distribució de les dades

Per tal de poder utilitzar propietats de distribucions (com la normal, etc...) en les dades, representem els histogrames dels atributs. Es mostren a continuació exemples d'alguns dels histogrames generats.



(a) *PER_ Lebron* (b) *OBPM_ Lebron* (c) *OWS_ Lebron*

(d) *PER_ Jordan* (e) *OBPM_ Jordan* (f) *OWS_ Jordan*

(g) *PER_ Kobe* (h) *OBPM_ Kobe* (i) *OWS_ Kobe*

Figura 1: Histogrames dels atributs

S'observa dels histogrames que alguns atributs, sobretot pel jugador Lebron, segueixen distribucions normals, com es dedueix segons l'histograma *PER_ Lebron* i *OBPM_ Lebron*.

S'observa també, que la majoria d'atributs no segueixen una distribució normal (com es dedueix a l'histograma *OWS_ Jordan*).

IMPORTANT: Els histogrames han sigut fets amb les dades estandaritzades per així poder visualitzar de millor manera com estan distribuïdes les dades, ja que en alguns casos els valors eren molt dispars i es dificultava la interpretació del gràfic.

4.3 Correlació de les variables

Mostrem ara l'anàlisi de les correlacions dels atributs de la base de dades per a cadascun dels jugadors.

Mencionar que els càlculs de les correlacions han sigut amb els datasets estandaritzats.

Atribut	**Corr. Lebron**
TS%	0.729
DRB%	0.651
TRB%	0.645
OBPM	0.927
BPM	0.885

Taula 1: Atributs rellevants per Lebron James

Atribut	**Corr. Jordan**
TS%	0.820
STL%	0.617
OBPM	0.919
BPM	0.887

Taula 2: Atributs rellevants per Michael Jordan

Atribut	**Corr. Kobe**
TS%	0.612
USG%	0.614
OWS	0.668
OBPM	0.948
BPM	0.931
VORP	0.678

Taula 3: Atributs rellevants per Kobe Bryant

S'observa a la Taula 1 que els atributs que presenten una correlació més elevada amb l'objectiu escollit *PER* són: *TS%*, *DRB%*, *TRB%*, *OBPM* i *BPM*.

S'observa d'igual manera a la Taula 2 que els atributs que presenten una correlació més elevada amb *PER* són *TS%*, *STL%*, *OBPM* i *BPM*.

Per la Taula 3, s'observa que els atributs que presenten una correlació més elevada amb *PER* són *TS%*, *USG%*, *OWS*, *OBPM*, *BPM* i *VORP*.

Degut a això, es decideix estudiar la possibilitat de predir la variable objectiu *PER* amb atributs diferents per a cada jugador, i comparar-ho amb els valors d'mse i r2 resultants d'usar els mateixos atributs per a cadascun.

Per altra banda, cal recalcar que només es mostren els atributs que presenten una correlació major al 0.5 (els atributs rellevants).

4.3.1 Pairplot Lebron James

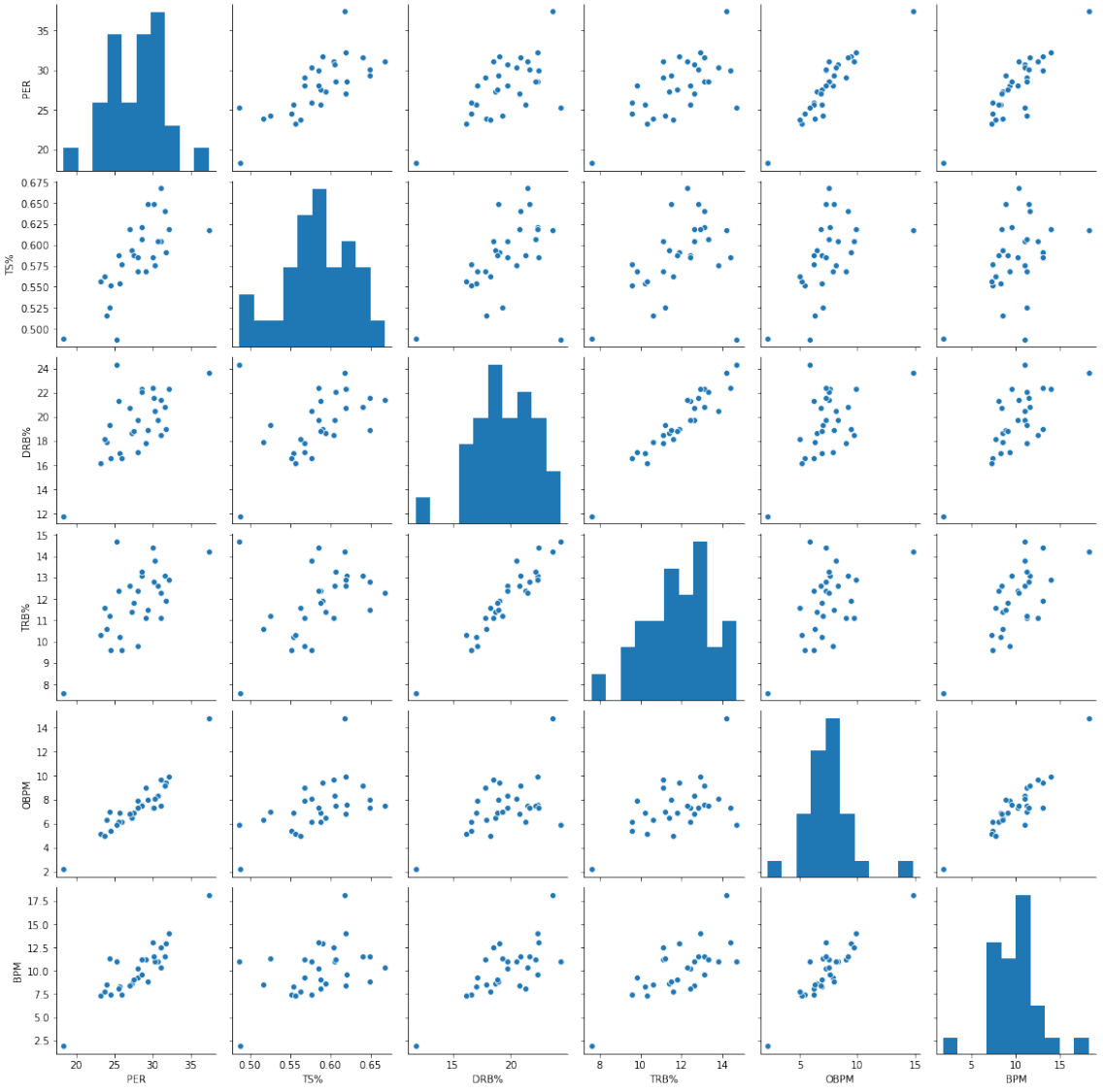


Figura 2: Pairplot dels atributs rellevants i l'objectiu

Els gràfics estan ordenats de la següent forma: PER, TS%, DRB%, TRB%, OBPM i BPM.

S'observa com els atributs semblen seguir distribucions normals. Per altra banda, observem que atributs com l'OBPM i el BPM semblen seguir una recta en relació a l'atribut objectiu PER.

Veiem també que la relació de PER amb atributs com DRB% i TRB% sembla no seguir cap tipus de distribució concreta, ja que presenta bastanta disperció.

4.3.2 Pairplot Michael Jordan

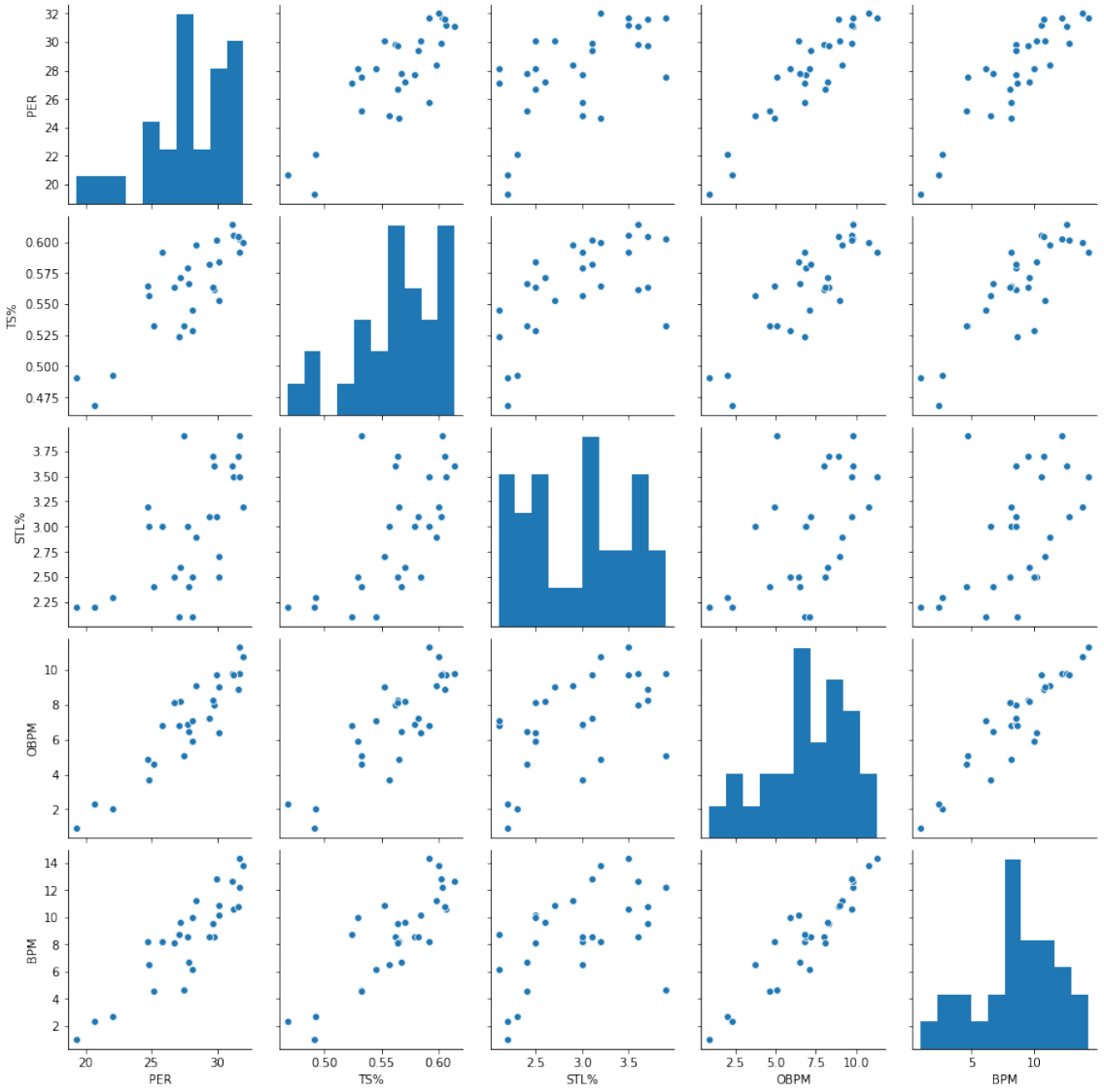


Figura 3: Pairplot dels atributs rellevants i l'objectiu

Els gràfics estan ordenats de la següent forma: PER, TS%, STL%, OBPM i BPM.

S'observa com els atributs no semblen seguir cap tipus de distribució concreta, tot i que PER i TS% podrien aproximar-se per una exponencial. Per altra banda, observem que atributs com l'OBPM i el BPM semblen seguir una recta ben marcada en relació a l'atribut objectiu PER, mentre que, tot i que la relació amb TS% també sembla ser lineal, s'observa més disperció que els casos anteriors, i pel que fa a STL%, no sembla seguir cap tipus de distribució concreta en relació a la variable objectiu.

4.3.3 Pairplot Kobe Bryant

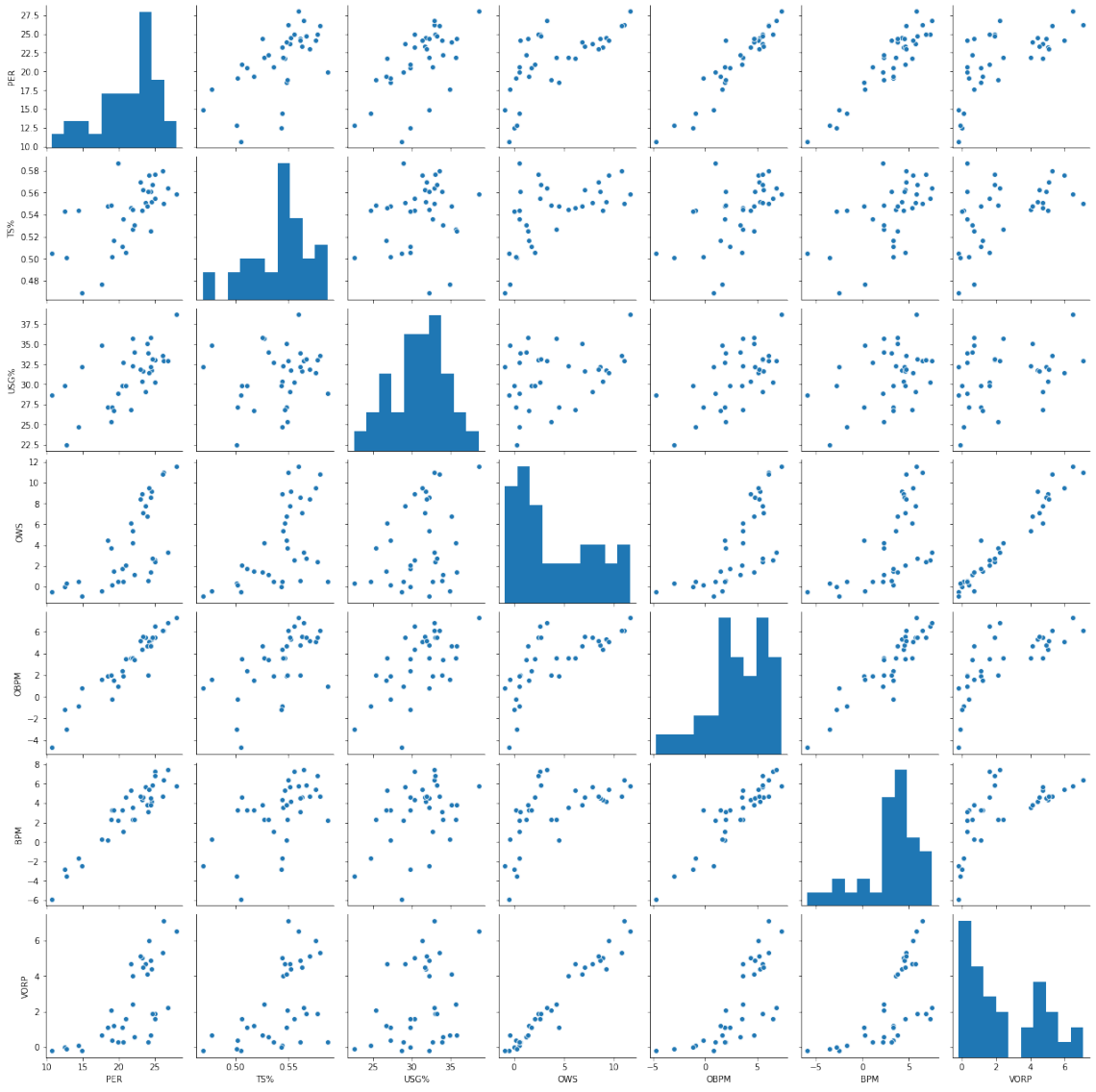


Figura 4: Pairplot dels atributs rellevants i l'objectiu

Els gràfics estan ordenats de la següent forma: PER, TS%, USG%, OWS, OBPM, BPM i VORP.

S'observa com l'atribut USG% sembla seguir una distribució normal, mentre que les PER, TS% i OBPM podrien aproximar-se per exponencials. Pel que fa a la resta d'atributs, no sembles seguir cap tipus de distribució concreta.

Pel que fa a la relació dels diversos atributs amb la variable objectiu PER, s'observa una relació aparentment lineal amb les variables OBPM i BPM, mentre que amb la resta de variables no presenten cap relació, ja que aparentment les dades estan bastant disperses.

Observem ara en un mapa de calor les correlacions entre els atributs més rellevants dels diversos jugadors trobats anteriorment. Es mostren a continuació, per ordre, els mapes de

calor dels jugadors Lebron James, Michael Jordan i Kobe Bryant, en aquest ordre.

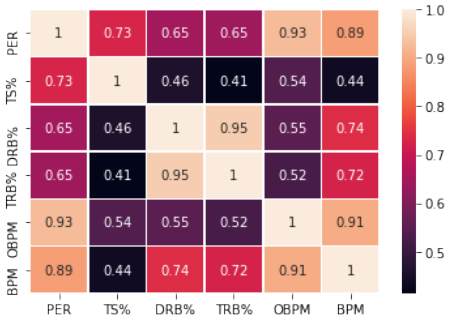


Figura 5: Correlacions Lebron James

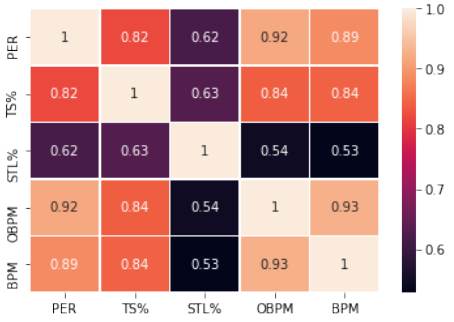


Figura 6: Correlacions Michael Jordan

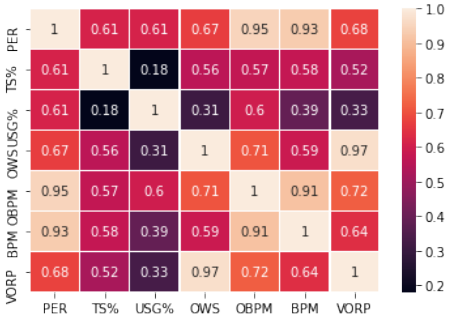


Figura 7: Correlacions Kobe Bryant

S'observa dels mapes de calor que existeix una forta correlació entre totes les dades rellevants.

A més, s'observa com el millor atribut (pel que fa a que té la millor correlació amb l'objectiu) és el *OBPM* en els 3 casos, que presenta una correlació de casi 1 en tots els casos.

Destaca el següent:

1. Pel que fa a la figura 5, observem que els atributs que presenten una major correlació amb l'objectiu, *PER* són *OBPM* i *BPM*, que presenten també una alta correlació entre ells.

2. Pel que fa a la figura 6, els atributs amb una correlació major o igual a 0.7 són *OBPM*, *BPM* i *TS%*. Destaca la elevada correlació entre els dos primers, de *0.93*, i tot i que la correlació entre els dos primers i el tercer no és tan elevada, també és molt significativa, sent de 0.84 en ambdós casos.

3. Pel que fa a la figura 7, els únics atributs que presenten una correlació amb l'objectiu major a 0.7 són *OBPM* i *BPM*.

Concloem l'estudi de les correlacions crivant les dades a les següents:

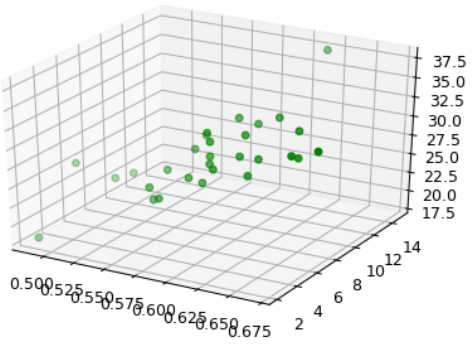
- Pel dataset Lebron James: *TS%*, *DRB%*, *TRB%*, *OBPM* i *BPM*.
- Pel dataset Michael Jordan: *TS%*, *STL%*, *OBPM* i *BPM*.
- Pel dataset Kobe Bryant: *TS%*, *USG%*, *OWS*, *OBPM*, *BPM* i *VORP*.

4.4 Representació de les dades

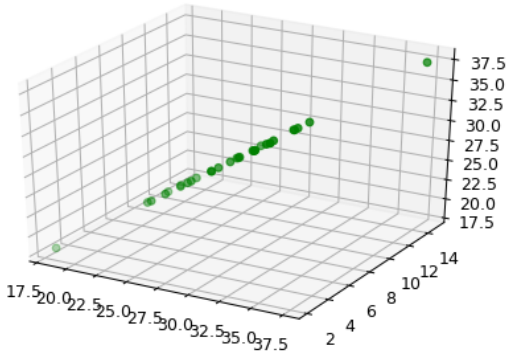
Decidim representar les dades del *PER* respecte dels atributs rellevants en R3 per veure si hi existeix alguna relació visual amb el valor del *PER*.

Es mostrarà a continuació, per cada jugador, un exemple dels gràfics generats al Notebook adjuntat amb l'estudi del Dataset, un d'ells amb relacions aparents i un on no s'observi cap relació entre les variables.

4.4.1 Representació 2 a 2 en R3 Lebron James

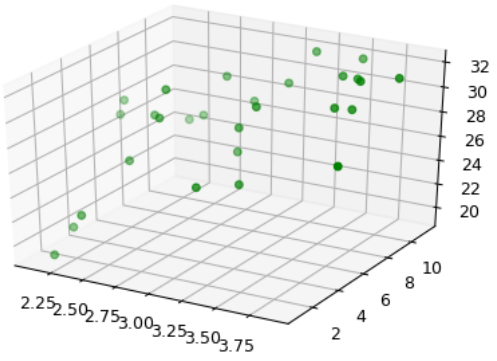


(a) USG% & OWS

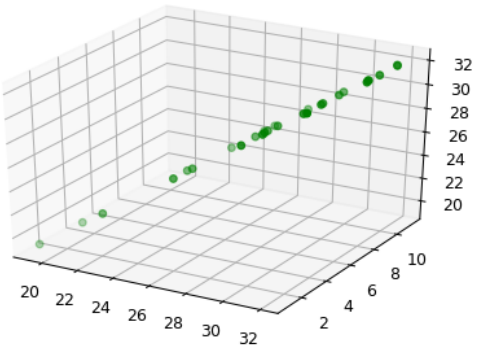


(b) TS% & OWS

4.4.2 Representació 2 a 2 en R3 Michael Jordan

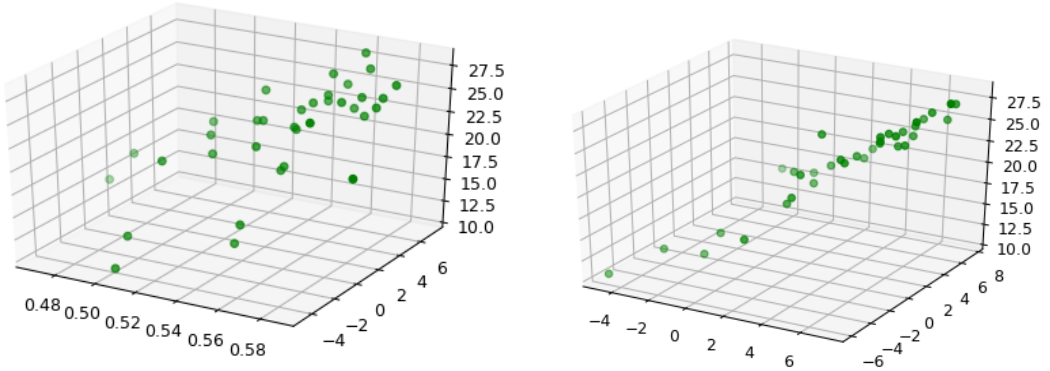


(a) OBPM & BPM



(b) TS% & BPM

4.4.3 Representació 2 a 2 en R3 Kobe Bryant



(a) USG% & OWS

(b) VORP & OWS

S'observa doncs com sembla que només certs atributs obtenen alguna relació (lineal) amb combinació dels altres atributs amb el PER. Es conclou que es pot predir el PER només usant 2 atributs dels que presenten una relació entre ells i la variable objectiu, com és el cas de:

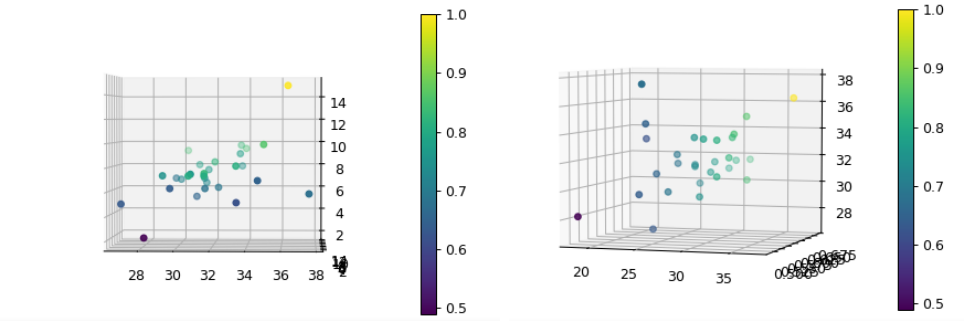
- TS% i OWS, per Lebron James
- TS% i BPM, per Michael Jordan
- VORP i OWS, per Kobe Bryant

Destacar que aquests només són alguns exemples de les combinacions que s'han trobat que presenten relacions entre elles. Més endavant s'estudiarà quina és la millor combinació d'atributs per tal de fer les prediccions.
Es procedeix a estudiar el comportament però en comptes de 2 a 2 fent 3 a 3.

4.4.4 Representació 3 a 3 en R4

Per a poder representar les dades amb 4 atributs (el PER i 3 més dels rellevants), s'ha decidit posar el valor de PER com el color del punt. Cada eix representa un dels atributs. Els atributs estan estandaritzats.
Una mostra dels resultats obtinguts són:

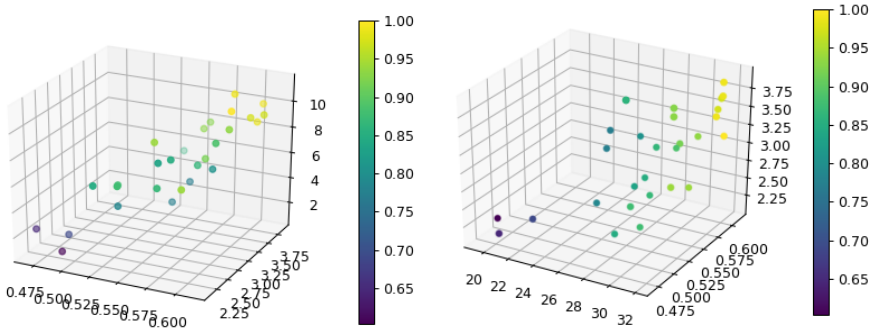
4.4.5 Representació 3 a 3 en R4 Lebron James



(a) OBPM & OWS & BPM

(b) OBPM & TS% & USG%

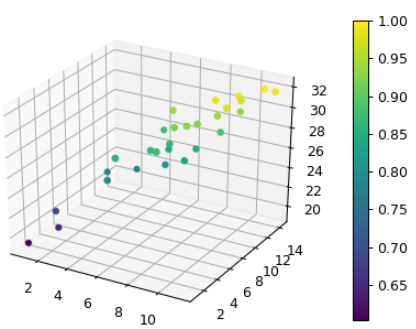
4.4.6 Representació 3 a 3 en R4 Michael Jordan



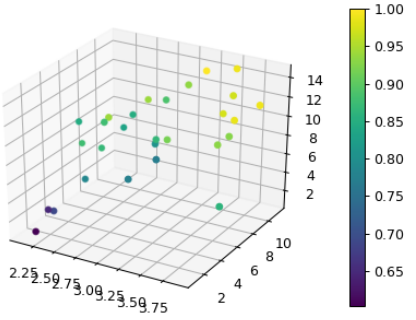
(a) OBPM & BPM & STL%

(b) TS% & OBPM & STL%

4.4.7 Representació 3 a 3 en R4 Kobe Bryant



(a) TS% & OWS & BPM



(b) OBPM & OWS & BPM

Dels gràfics ens adonem que no hi existeix cap relació adicional a les ja trobades amb les representacions del *PER* fetes anteriorment, i per això deduïm que no hi haurà una gran millora dels resultats una vegada afegim més atributs per al nostre regressor. Igualment s'estudia si la millora que obtenim és considerable o si pel contrari amb un nombre menor de variables s'obté un resultat prou bo.

4.5 Regressió lineal

Per a trobar el millor regresor lineal estudiem, per als atributs rellevants, el seu MSE i el R2 score.

S'utilitzen els resultats obtinguts per comparar-los amb altres regresors lineals.
Per al regresor lineal simple, s'utilitza el mètode *LinearRegressor()* de la llibreria sklearn[2].

IMPORTANT: les dades obtingudes a continuació són mitjançant tots els valors dels que es disposa.

Lebron	**Original**		**Estandaritzat**	
	MSE	R2	MSE	R2
TS%	6.12	0.53	18.97	0.45
DRB%	7.53	0.42	17.53	−2.34
TRB%	7.62	0.41	16.41	−1.06
OBPM	1.83	0.85	3.66	0.84
BPM	2.82	0.78	2.47	0.65

Taula 4

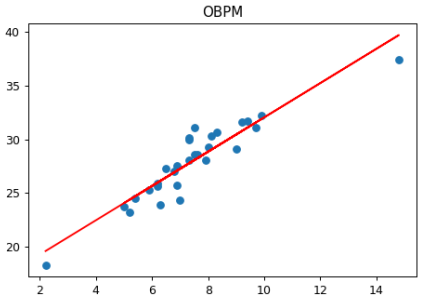


Figura 8: Regressor de OBPM

S'observa com, per a tots els atributs menys *BPM*, el MSE augmenta quan les dades són estandaritzades, i que el R2 score disminueix significativament en alguns casos, per tant, es decideix utilitzar per a tots els regresors les dades sense estandaritzar.

Es conclou doncs que el millor regressor lineal simple per predir el *PER* de Lebron James és l'*OMPB*, en tenir l'r2 més elevat i el menor mse.

El valor obtingut del PER és: 27.93793103448275.

Jordan	**Original**		**Estandaritzat**	
	MSE	R2	MSE	R2
TS%	3.46	0.67	2.00	0.84
STL%	6.57	0.38	8.52	−0.05
OBPM	1.63	0.84	2.03	0.65
BPM	2.63	0.78	1.42	0.65

Taula 5

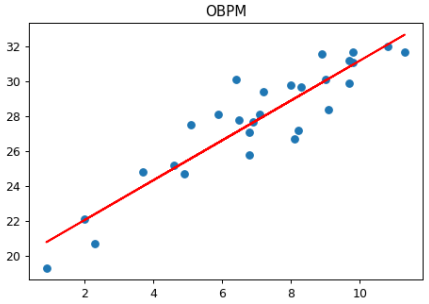


Figura 9: Regressor de OBPM

S'observa com, per als atributs *TS%* i *BMP*, el MSE disminueix quan les dades són estandaritzades, mentre que per la resta augmenta. Per altra banda, s'observa que el R2 score disminueix significativament en alguns casos, per tant, es decideix utilitzar per a tots els regresors les dades **sense estandaritzar**.

[2]La informació sobre les llibreries i funcions utilitzades es troba a 2.1

Es conclou doncs que el millor regressor lineal simple per predir el *PER* de Michael Jordan és l'*OMPB*, en tenir l'r2 més elevat i el menor mse.

El valor obtingut del PER és: 27.839285714285715.

Kobe	**Original**		**Estandaritzat**	
	MSE	R2	MSE	R2
TS%	11.08	0.37	20.30	0.23
USG%	11.03	0.37	14.27	0.11
OBPM	1.77	0.90	1.16	0.86
BPM	2.33	0.86	2.84	0.83
OWS	9.81	0.44	10.74	0.41
VORP	9.58	0.46	9.58	0.53

Taula 6

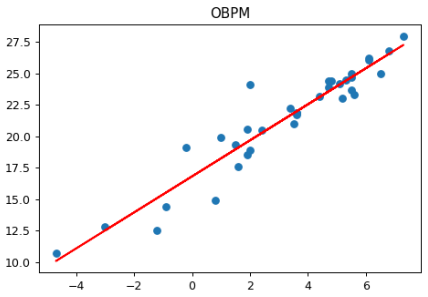


Figura 10: Regressor de OBPM

S'observa com, per a tots els atributs excepte *OBPM*, el MSE augmenta quan les dades són estandaritzades, en alguns casos molt significativament. Per altra banda, s'observa que el R2 score disminueix en tots els casos excepte per l'atribut *VORP*, per tant, es decideix utilitzar per a tots els regresors les dades sense estandaritzar.

Es conclou doncs que el millor regressor lineal simple per predir el *PER* de Kobe Bryant és l'*OMPB*, en tenir l'r2 més elevat i el menor mse.

El valor obtingut del PER és: 21.397142857142857.

RECALCAR: Pels 3 jugadors professionals, el millor regressor lineal trobat és el que prediu el *PER* usant l'atribut *OBPM*.

4.6 Regressió multilineal amb 3 atributs

Per tal d'incrementar la complexitat del model i veure si millora la precissió en la predicció de resultats, es planteja un model de regressió multilineal.

El primer apropament a aquest model planteja l'ús dels 3 atributs que són rellevants pels 3 jugadors a la vegada: TS%, OBPM i BPM.

Els resultats obtinguts, per cada jugador, són:

Jugador	**MSE**	**R2**	**Predicció**
Lebron James	4.02	−0.06	27.93793103448275
Michael Jordan	2.29	0.77	27.839285714285715
Kobe Bryant	3.62	0.84	21.397142857142853

Taula 7

S'observa que, en comparació al regressor lineal estudiat a l'apartat 4.5, per a tots els jugadors l'MSE augmenta, i l'R2 disminueix, de manera que la precissió del model ha disminuit.

Tot i així, es seguirà estudiant la possibilitat d'usar un mètode de regressió multilineal que, usant altres combinacions d'atributs rellevants, millori la predicció de l'objectiu.

Destacar que tant el regressor lineal simple com el regressor multilineal explicat determinen que el millor jugador és Lebron James, seguit amb molt poca diferència per Michael Jordan i, amb més distància, Kobe Bryant.

4.7 Regressió multilineal

Es procedeix ara a intentar millorar la precisió del regresor afegint-li major complexitat. S'analitzen ara totes les possibles combinacions lineals del regressor multilineal simple de la llibreria sklearn (*LinearRegressor()*) amb totes les variables rellevants[3].

Els resultats més significatius obtinguts, són els següents:

Jugador		Atributs	Valor
Lebron	Millor R2	('TS%', 'OBPM', 'BPM')	0.91
Lebron	Millor MSE	('TS%', 'OBPM', 'BPM')	0.92
Jordan	Millor R2	('OBPM', 'BPM')	0.71
Jordan	Millor MSE	('STL%', 'OBPM')	1.92
Kobe	Millor R2	('USG%', 'BPM', 'OWS', 'VORP')	0.96
Kobe	Millor MSE	('USG%', 'BPM', 'OWS', 'VORP')	0.66

Taula 8: Millors valors de MSE i R2 obtinuts en la regressió multilineal

IMPORTANT: Les combinacions d'atributs tant per aquest model com pels que s'estudien a continuació s'han cercat dins la tupla d'atributs més rellevants per a cada jugador:

- ("TS%", "DRB%", "TRB%", "OBPM", "BPM"), per Lebron James.
- ("TS%", "STL%", "OBPM", "BPM"), per Michael Jordan.
- ("TS%", ÜSG%", "OBPM", "BPM", "OWS", "VORP"), per Kobe Bryant.

Abans d'explicar els resultats cal comentar que els estudis dels regresors multilineals s'han dut a terme amb els datatset separat en train i test i que s'han repetit els càlculs 100 vegades per cada conjunt d'atributs trobats per a evitar valors sobre o infraestimats deguts a fenòmens aleatoris i evitar també possibles overfittings del model.

S'observa que, en el cas de Lebron James i Kobe Bryant, l'error MSE disminueix respecte el millor model trobat fins el moment, mentre que l'R2 augmenta. Deduïm, per tant, que el regressor multilineal trobat és el model que millor prediu la variable objectiu PER per aquests dos jugadors.

Pel que fa a Michael Jordan, s'observa que l'error MSE augmenta en 0.3, i l'R2 disminueix en 0.1, pel que el millor model per predir la variable objectiu PER segueix sent una regressió lineal usant l'atribut OBPM.

[3]Per a fer aquestes combinacions s'ha utilitzat el mètode *combinations*

4.8 Regressió multilineal amb Lasso

Estudiem ara els millors valors assolits de totes les combinacions d'atributs rellevants mitjançant el mètode de regressió multilineal *Lasso()*[4].

Els resultats més significatius obtinguts són els següents:

Jugador		Atributs	Valor
Lebron	Millor R2	('TS%', 'TRB%', 'OBPM')	0.80
Lebron	Millor MSE	('TS%', 'TRB%', 'OBPM')	1.82
Jordan	Millor R2	('STL%', 'OBPM')	0.75
Jordan	Millor MSE	('TS%', 'STL%', 'OBPM')	1.74
Kobe	Millor R2	('TS%', 'USG%', 'BPM', 'OWS', 'VORP')	0.94
Kobe	Millor MSE	('USG%', 'OBPM', 'BPM', 'OWS', 'VORP')	0.88

Taula 9: Millors valors de MSE i R2 obtinuts en la regressió Lasso

S'observa com el regressor multilineal amb el mètode *Lasso* no millora ni en R2 ni en MSE al regressor multilineal simple en els casos de Lebron James i Kobe Bryant, mentre que en el cas de Michael Jordan hi ha una certa millora, tot i que segueix sent pitjor que la predicció del model lineal simple.

Destacar que els valors de les prediccions són 28.22013235806344, 28.120490339285748 i 21.61327561327562, segons l'ordre mostrat a la taula.

4.9 Regressió multilineal amb BayessianRidge

Estudiem ara els millors valors obtinguts de totes les combinacions d'atributs rellevants mitjançant el mètode de regressió multilineal *BayessianRidge()*[5].

Els resultats més significatius obtinguts són els següents:

Jugador		Atributs	Valor
Lebron	Millor R2	('TS%', 'OBPM')	0.83
Lebron	Millor MSE	('TS%', 'OBPM')	1.68
Jordan	Millor R2	('TS%', 'STL%', 'OBPM')	0.77
Jordan	Millor MSE	('TS%', 'STL%', 'OBPM')	1.82
Kobe	Millor R2	('TS%', 'USG%', 'BPM', 'OWS', 'VORP')	0.96
Kobe	Millor MSE	('TS%', 'USG%', 'BPM', 'OWS', 'VORP')	0.66

Taula 10: Millors valors de MSE i R2 obtinuts en la regressió BayessianRidge

S'observa com el regresor multilineal amb mètode Lasso millora l'R2 score pel que fa a la predicció del PER per a Kobe Bryant, pel que es situa com a millor model fins el moment per a aquest jugador.

Destacar que els valors de les prediccions són 28.220104137931084, 28.120490620490372 i 21.61327561327562, segons l'ordre mostrat a la taula.

[4]Usant el valor del paràmetre $a = 0.01$

[5]Usant el valor del paràmetre $t = 10^{-12}$

4.10 Regressió multilineal amb ElasticNet

Estudiem ara els millors valors obtinguts de totes les combinacions d'atributs rellevants mitjançant el mètode de regressió multilineal *ElasticNet()*[6].

Els resultats més significatius obtinguts són els següents:

Jugador		Atributs	Valor
Lebron	Millor R2	('TS%', 'TRB%', 'OBPM')	0.78
Lebron	Millor MSE	('TRB%', 'OBPM', 'BPM')	2.20
Jordan	Millor R2	('STL%', 'OBPM')	0.75
Jordan	Millor MSE	('STL%', 'OBPM', 'BPM')	2.07
Kobe	Millor R2	('USG%', 'BPM', 'OWS')	0.93
Kobe	Millor MSE	('TS%', 'USG%', 'OBPM', 'BPM', 'OWS', 'VORP')	1.19

Taula 11: Millors valors de MSE i R2 obtinuts en la regressió ElasticNet

S'observa com el regresor multilineal amb mètode *ElasticNet* no millora ni en R2 ni en MSE cap dels regressors estudiats fins el moment, de manera que tindrà una precissió menys fiable que la resta de models estudiats fins el moment.

Destacar que els valors de les prediccions són 28.22013235806344, 28.120490339285748 i 21.61327561327562, segons l'ordre mostrat a la taula.

4.11 Regressió polinòmica

Degut a que fins ara els millors models han resultat el model lineal simple i el model multilineal simple, es decideix estudiar la introducció de més complexitat al model usant una regressió polinòmica amb un únic atribut.

Per tal de trobar la millor regressió polinòmica possible, per cada jugador s'estudia quin dels atributs més rellevants proporciona un valor d'error MSE més baix i un d'R2 més elevat, amb exponents d'entre 2 i 5.

Els resultats més significatius obtinguts són els següents:

Jugador		Atributs	Valor
Lebron	Millor R2	('OBPM', grau: 4)	0.89
Lebron	Millor MSE	('OBPM', grau: 4)	1.36
Jordan	Millor R2	('OBPM', grau: 4)	0.86
Jordan	Millor MSE	('OBPM', grau: 4)	1.38
Kobe	Millor R2	('OBPM', grau: 4)	0.90
Kobe	Millor MSE	('OBPM', grau: 4)	1.68

Taula 12: Millors valors de MSE i R2 obtinuts en la regressió Polinòmica

S'observa com el regresor polinòmic lineal no millora ni en R2 ni en MSE cap dels regressors estudiats fins el moment, ja que els valors del MSE i del R2 són majors i menors,

[6]Usant el valor del paràmetre $t = 10^{-6}$

respectivament, dels millors models trobats fins el moment (multilineal simple per Lebron James i BayessianRidge per Kobe Bryant). Pel que fa a Michael Jordan, observem que disminueix l'error MSE i augmenta l'error R2, pel que deduïm que usar el model polinòmic de grau 4 prediu millor l'objectiu que usar el model lineal simple.

Destacar que els valors de les prediccions són 27.937931034482805, 27.83928571428571 i 21.397142857142857, segons l'ordre mostrat a la taula.

4.12 Versió final del regresor i interval de confiança

Per finalitzar l'estudi dels regresors, es procedeix a estudiar quin és el marge d'error d'alguns dels models explicats anteriorment, generant el seu interval de confiança del 75%. **IMPORTANT**: En cas de mostrar-se dues imatges, la de l'esquerra correspon al CrossValidation que proporciona l'R2 més elevat i la de la dreta a l'MSE.

4.12.1 CrossValidation model lineal

Es mostra a continuació l'interval de confiança del 75% pels atributs que donen una millor precissió per cadascun dels jugadors:

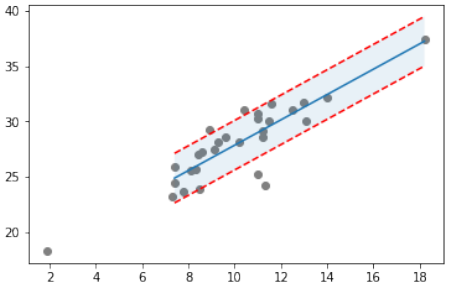


Figura 11: Lebron James

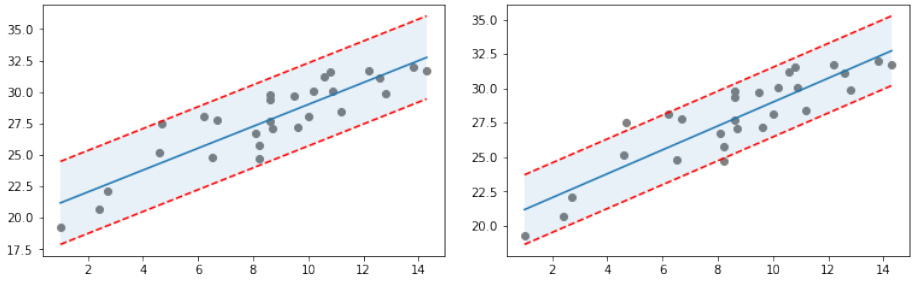


Figura 12: Michael Jordan

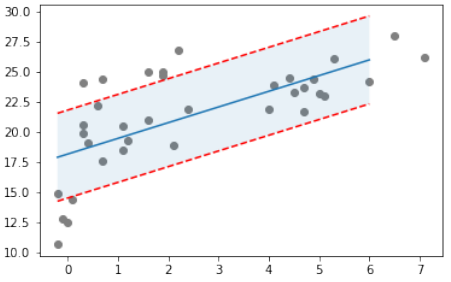


Figura 13: Kobe Bryant

Els resultats obtinguts són els següents:

- Per Lebron James, Figura 11, el millor regressor lineal és usant l'atribut OBPM, amb el qual s'obté un MSE de 1.07 i un R2 de 0.93. Destacar que hi ha altres atributs amb valors d'MSE i R2 molts similars, pel que també es podrien usar els atributs TRB% o TS% per fer les prediccions.

- Per Michael Jordan, Figura 12, el millor regressor lineal pel que fa a l'MSE és usant l'atribut TS% (1.20), i pel que fa a l'R2 usant l'atribut STL% (0.87). Mitjançant el CrossValidation s'ha observat una millora tant en l'error MSE com en l'R2, en comparació a l'estudi realitzat a l'apartat 4.5.

- Per Kobe Bryant, Figura 13, el millor regressor lineal és usant l'atribut USG%, que dona els valors 6.73 i 0.57 d'MSE i R2, respectivament. S'observa que en aquest cas els valors obtinguts dels errors indiquen que la precissió del model no és elevada, ja que l'objectiu és obtenir un R2 proper a 1 i un MSE baix.

4.12.2 CrossValidation model Lasso

Es mostra a continuació l'interval de confiança del 75% pels atributs que donen una millor precissió per cadascun dels jugadors:

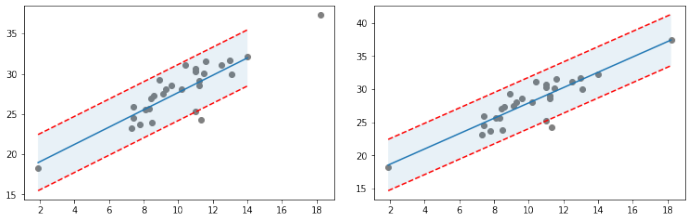


Figura 14: Lebron James

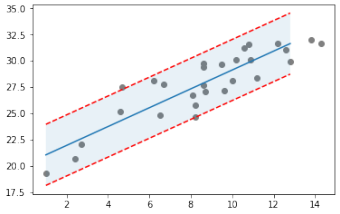


Figura 15: Michael Jordan

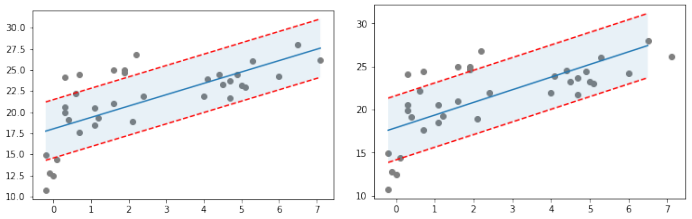


Figura 16: Kobe Bryant

Els resultats obtinguts són els següents:

- Per Lebron James, Figura 14, el millor regressor lasso pel que fa al R2 és usant la combinació d'atribut TS%, TRB% i BPM, amb un valor de 0.91. En canvi, pel que fa al MSE, la millor combinació és DRB%, OBPM i BPM, amb un valor de 0.78.
- Per Michael Jordan, Figura 15, el millor regressor lasso és usant els atributs TS% i STL%, amb uns valors de MSE i R2 de 0.79 i 0.90, respectivament.
- Per Kobe Bryant, Figura 16, el millor regressor lasso pel que fa al R2 és usant la combinació d'atributs TS%, OBPM, BPM i VORP, amb un valor de 0.64. En canvi, pel que fa al MSE, la millor combinació és TS%, USG%, BPM i VORP, amb un valor de 4.53.

4.12.3 CrossValidation model BayessianRidge

Es mostra a continuació l'interval de confiança del 75% pels atributs que donen una millor precissió per cadascun dels jugadors:

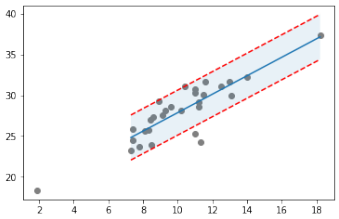


Figura 17: Lebron James

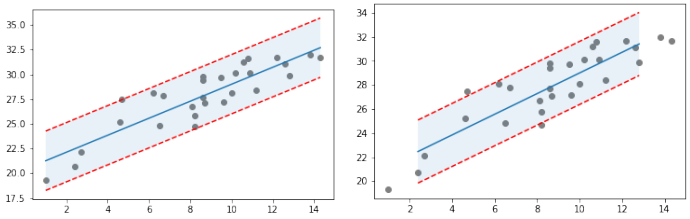


Figura 18: Michael Jordan

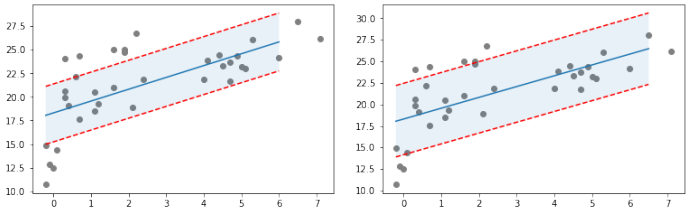


Figura 19: Kobe Bryant

Els resultats obtinguts són els següents:

- Per Lebron James, Figura 17, el millor regressor BayessianRidge és usant la combinació d'atributs TS% i DRB%, que donen els valors d'MSE i R2: 0.57 i 0.94, respectivament.
- Per Michael Jordan, Figura 18, el millor regressor BayessianRidge segons l'MSE és usant la combinació d'atributs TS% i STL%, que donen un valor de 0.97, mentre que per l'R2, la millor combinació resulta TS%, OBPM i BPM, resultant en un valor de 0.90.
- Per Kobe Bryant, Figura 19, el millor regressor BayessianRidge pel que fa al R2 és usant la combinació d'atributs USG%, BPM i OWS, amb un valor de 0.66. En canvi, pel que fa al MSE, la millor combinació és TS%, USG% OBPM, BPM i VORP, amb un valor de 3.26.

4.12.4 CrossValidation model ElasticNet

Es mostra a continuació l'interval de confiança del 75% pels atributs que donen una millor precissió per cadascun dels jugadors:

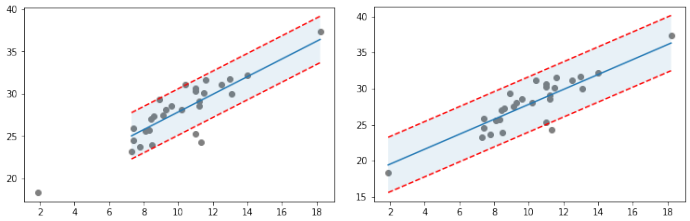


Figura 20: Lebron James

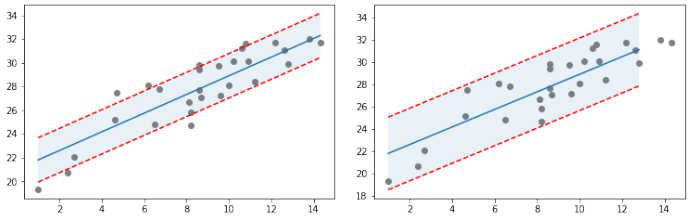


Figura 21: Michael Jordan

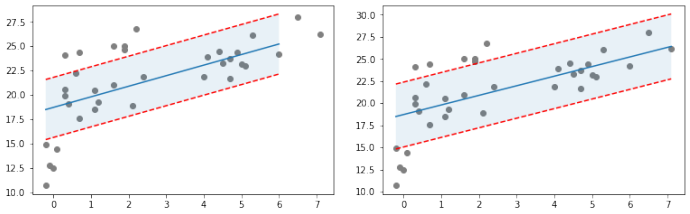


Figura 22: Kobe Bryant

Els resultats obtinguts són els següents:

- Per Lebron James, Figura 20,el millor regressor ElasticNet segons l'R2 és usant la combinació d'atributs TS%, OBPM I BPM, que dona un valor de 0.91, mentre que per l'MSE, la millor combinació resulta DRB%, TRB% i OBPM, resultant en un valor de 0.81.
- Per Michael Jordan, Figura 21, el millor regressor ElasticNet segons l'MSE és usant la combinació d'atributs OBPM i STL%, que donen un valor de 1.04, mentre que per l'R2, la millor combinació resulta OBPM i BPM, resultant en un valor de0.84.
- Per Kobe Bryant, Figura 22, el millor regressor ElasticNet pel que fa al MSE és usant la combinació d'atributs TS%, USG%, OBPM, BPM i VORP, amb un valor de 3.52. En canvi, pel que fa al R2, la millor combinació és USG% i VORP, amb un valor de 0.67.

4.12.5 CrossValidation model multilineal

Es mostra a continuació l'interval de confiança del 75% pels atributs que donen una millor precissió per cadascun dels jugadors:

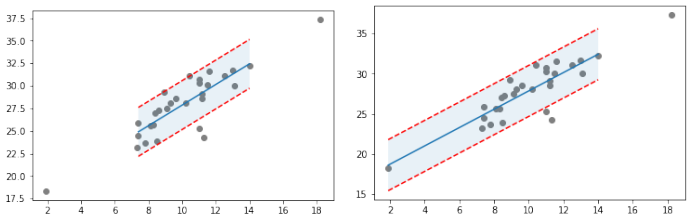


Figura 23: Lebron James

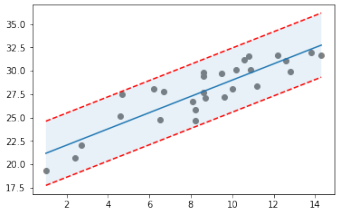


Figura 24: Michael Jordan

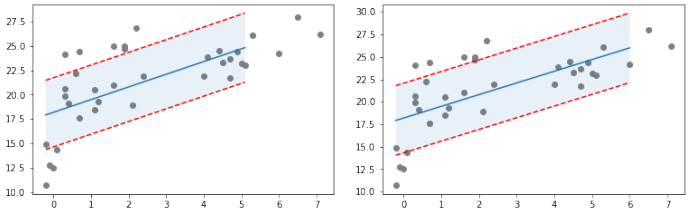


Figura 25: Kobe Bryant

Els resultats obtinguts són els següents:

- Per Lebron James, Figura 23,el millor regressor multilineal segons l'R2 és usant la combinació d'atributs TS%, OBPM i DRB%, que dona un valor de 0.94, mentre que per l'MSE, la millor combinació resulta TS%, TRB%, OBPM i BPM, resultant en un valor de 0.71.
- Per Michael Jordan, Figura 24, el millor regressor multilineal és usant la combinació d'atributs OBPM i BPM, que proporcionen uns valors de R2 i MSE de 0.83 i 1.43, respectivament.
- Per Kobe Bryant, Figura 25, el millor regressor multilineal pel que fa al R2 és usant la combinació d'atributs USG%, BPM i OBPM, amb un valor de 0.77. En canvi, pel que fa al MSE, la millor combinació és USG%, OWS i VORP, amb un valor de 2.06.

5 Cerca d'hiperparàmetres

Un cop realitzat l'estudi fet anteriorment a la secció 4.12, s'estudia quin dels models (un cop aplicat el CrossValidation) dona millors resultats per a cada jugador.

A continuació, es mostra una taula amb el millor mètode de regressió per trobar el *PER* per a cada jugador, respecte l'MSE i l'R2, juntament amb els errors que genera:

Jugador	Millor regressor	MSE	R2
Lebron James	BayessianRidge (TS% i DRB%)	0.57	0.94
Michael Jordan	Lasso (TS% i STL%)	0.79	0.90
Kobe Bryant	Multilineal (USG%, OWS i VORP)	2.06	0.53
Kobe Bryant	Multilineal (USG%, OBPM i BPM)	3.66	0.78

Taula 13: Regressors amb millor relació MSE-R2 trobats fins el moment

Es decideix d'aquesta manera realitzar una cerca d'hiperparàmetres per tal d'intentar incrementar la precissió dels models BayessianRidge i Lasso.

5.1 Lasso

Paràmetres	**Valors**
alpha	0
	⋮
	1.81
	⋮
	10
selection	cyclic
	random
tol	10^{-12}
	⋮
	0.00017
	⋮
	10^{-1}

Figura 26: Hiperparàmetres model Lasso

Per a obtenir la millor precisió possible amb el Lasso s'ha proposat trobar la millor combinació de paràmetres usant la funció *RandomizedSearchCV* de la llibreria *sklearn*.

Els paràmetres i valors testejats es mostren a la taula del costat.

Els resultats de la cerca dels millors hiperparàmetres són els valors remarcats a la taula.

Destacar que, després de realitzar diverses proves, s'ha comprovar que usant els paràmetres trobats no s'aconsegueix millorar de forma significativa la precissió del model.

5.2 BayessianRidge

Paràmetres	Valors
n_iter	0 ⋮ 275 ⋮ 500
tol	10^{-12} ⋮ 0.020 ⋮ 10^{-1}
alpha_1	10^{-12} ⋮ 0.024 ⋮ 10^{-1}
alpha_2	10^{-12} ⋮ 0.008 ⋮ 10^{-1}
lambda_1	10^{-12} ⋮ 0.050 ⋮ 10^{-1}
lambda_2	10^{-12} ⋮ 0.090 ⋮ 10^{-1}

Figura 27: Hiperparàmetres BayesianRidge

Per a obtenir la millor precisió possible amb el BayessianRidge s'ha proposat trobar la millor combinació de paràmetres usant la funció *RandomizedSearchCV* de la llibreria *sklearn*.

Els paràmetres i valors testejats es mostren a la taula del costat.

Els resultats de la cerca dels millors hiperparàmetres són els valors remarcats a la taula.

Destacar que, després de realitzar diverses proves, s'ha comprovar que usant els paràmetres trobats no s'aconsegueix millorar de forma significativa la precissió del model.

6 Anàlisi i Conclusions

Resumim els resultats obtinguts del nostre estudi:

- En definir el millor jugador com aquell que té més capacitats i aporta més al joc en general, no en un talent concret, el millor jugador serà aquell que tingui la PER més alta (*Player efficiency rating*). D'aquesta manera, el treball s'ha centrat en predir la PER de cada jugador mitjançant l'estudi de diversos regressors.
- Els atributs que més s'han usat per fer les prediccions i han destacat donant resultats han estat TS% (Percentatge de tirs), OBPM (*Box* ofensiva) i BPM (*Box Plus/Minus*), pel que per tal de ser considerat un bo en l'esport, aquestes són les àrees on més s'hauria de centrar a millorar un jugador.
- La cerca d'hiperparàmetres pels mètodes Lasso i BayessianRidge no milloren de forma significativa els valors d'error de l'MSEi de l'R2 score.
- Els millor regressors trobats són:
  - BayessianRidge usant TS% i DRB%, per Lebron James, amb un R2 de 0.94.
  - Lasso usant TS% i STL%, per Michael Jordan, amb un R2 de 0.90.
  - Regressió lineal usant OBPM, per Kobe Bryant amb un R2 de 0.90.
- Els valors predits per cada jugador són:
  - 28.95927462025407
  - 27.93157894736842
  - 21.68040897432857

A partir d'aquestes dades concluïm que els millors jugadors són Lebron James en primer lloc, amb poca diferència respecte de Michael Jordan, que ocupa la segona posició, i per últim, Kobe Bryant.

Cal destacar el fet que s'ha fet aquest podi amb un anàlisi de l'eficiència general del jugador, per tant, és possible que un jugador no catalogat com el millor sigui més bo en un sector concret com podria ser el percentatge de bloquejos.

Sabent que un valor d'1 en l'R2 score indica un ajust perfecte i, per tant, un model molt fiable per a predir, mentre que un valor de 0 indicaria que el model no aconsegueix ajustar les dades en absolut, assegurem que el nostre regresor funciona de manera correcta observant l'R2 Score màxim aconseguit per a cada jugador, mencionats anteriorment.