EAD: Estudio sobre NBA dataset y aplicación de PCA y K-Means

Álvaro Sellart Álvarez

11 de enero de 2018

Índice

1. Introducción

En el presente trabajo se pretende aplicar técnicas vistas en la asignatura de *Exploración y Análisis de Datos* (EAD). El conjunto de datos seleccionado contiene estadísticas sobre jugadores de la NBA, sobre dicho conjunto se pretende realizar un estudio inicial de los datos, aplicar *Principal Component Analysis* (PCA) y técnicas de *clustering* (K-Means).

Este documento pretende resumir el trabajo realizado, primero se presenta el análisis de los datos realizado describiendo el conjunto y resaltando los aspectos más relevantes. Posteriormente se detalla la aplicación de PCA mostrando y examinando los resultados. Finalmente se presentan los resultados de aplicar K-Means. Se ha optado por no incluir todos los resultados calculados en el código y así mismo no se incluye ninguna parte del código implementado (está disponible en la entrega) para facilitar la lectura del documento.

2. Análisis de los datos

2.1. Resumen

Los datos pertenecen a un *dataset* de la página *Kaggle*[1] subido por el usuario *DrGuillermo*. El *dataset* se compone de dos conjuntos de datos, uno *Players.csv*, dedicado a la información de los jugadores (altura, peso, ciudad de nacimiento, etc) y otro, *Seasons_Stats.csv*, contiene estadísticas de cada jugador por temporada.

2.2. Descripción

Para facilitar la lectura del trabajo a continuación sólo se describen las variables que han sido consideradas de mayor interés de los dos conjuntos de datos. En la entrega del trabajo se incluye un archivo, *glosary.txt*, con las descripciones de todas las variables del conjunto de datos.

2.2.1. Archivo Players

Contiene información general de 3992 jugadores:

- **Player**, nombre y apellidos.
- **height**, altura (cm).
- **weight**, peso (kg).

2.2.2. Archivo Seasons_Stats

Contiene información de cada jugador por cada temporada. Siguiendo un criterio personal se ha hecho una división de las variables, por un lado las variables *tradicionales* son aquellas que históricamente han sido más usadas, por ejemplo: puntos, rebotes o asistencias totales. Y por otro lado las variables *avanzadas* que son más recientes, se extraen mediante fórmulas usando variables *tradicionales* y en la actualidad están siendo muy usadas para analizar el baloncesto.

Estadísticas tradicionales:
Estadísticas de uso más tradicional durante la historia, por ejemplo los puntos, asistencias o lanzamientos lanzados han sido recogidos desde las primeras planillas estadísticas. Otras en cambio se empezaron a recoger en diferentes años, minutos jugados (1951-52), tapones (1973-74), triples lanzados (1979-80). Estas variables son:

- **Year**, año de temporada.
- **Player**, nombre y apellidos del jugador.
- **G**, partidos jugados.
- **MP**, total minutos jugados.

[1] https://www.kaggle.com/drgilermo/nba-players-stats

- Estadística de lanzamientos lanzados (**FGA**, **X2PA**, **X3PA**, **FTA**), anotados (**FG**, **X2P**, **X3P**, **FT**) y su porcentaje (**FG.**, **X2P.**, **X3P.**, **FT.**). Siendo lanzamientos de campo (FG), tiros de dos puntos (X2P), triples (X3P), y tiros libres (FT).
- Rebotes ofensivos (**ORB**), defensivos (**DRB**) y totales (**TRB**).
- **AST**, **STL**, **BLK**, **TOV**, **PTS**, total de asistencias, robos, tapones, perdidas y puntos anotados.

Estadísticas avanzadas:
Intentan ponderar y contextualizar las estadísticas tradicionales mediante formulas[2]. Algunos ejemplos:

- **PER** (Player Efficiency Rating), evalúa el rendimiento general del jugador sumando estadísticas positivas y restando las negativas.
- **TRB %** (Total Rebound Percentage), estimación del porcentaje de rebotes capturados por un jugador entre los totales disponibles cuando éste estaba en juego.
- **Usg %** (Usage Percentage), estima el porcentaje de jugadas en ataque que son acabadas por un jugador (lanzando a canasta, perdiendo el balón o sacando una falta personal).

2.3. Estudio de los datos

El conjunto recoge 24624 instancias con 50 variables, 47 de ellas numéricas y 3 categóricas. Las variables numéricas son de dos tipos, por un lado hay variables de números enteros que recogen los totales (puntos, rebotes asistencias...) y por otro lado existen variables que se sitúan en el rango de [0,1] como los porcentajes de tiro o las estadísticas *avanzadas*. Se han calculado las medidas de centralidad y dispersión para todas las variables, están disponibles ejecutando el script *EAD_ Analisis.R*, pero al ser un número alto de variables y no existir aspectos destacables se han omitido en este documento. A continuación se muestran los aspectos más importantes del estudio de los datos.

2.3.1. Posiciones

Una de las variables más interesantes es la posición del jugador, ya que sobre esta variable se basa el estudio del clustering, analizando si *K-Means* agrupa los jugadores con la posición original o si propone otras agrupaciones de jugadores.

Históricamente han existido cinco posiciones diferentes: Base (Point Guard), Escolta (Shooting Guard), Alero (Small Forward), Ala-Pivot (Power Forward) y Pivot (Center). Es importante considerar que si bien la asignación de posiciones trata de agrupar los jugadores según el estilo de juego y posición en el campo, muchas veces esta clasificación es ambigua y subjetiva. En el dataset original existen casos donde a los jugadores se les asignan dos posiciones, para facilitar el estudio se han modificado estas posiciones ambiguas transformándolas sólo a una posición. Por ejemplo, si un jugador es Point Guard y Shooting Guard se transforma únicamente a Shooting Guard. También existen posiciones usadas en los primeros años que actualmente tienen otro nombre (Guard y Forward). Una vez realizadas las modificaciones la distribución de posiciones resulta muy balanceada:

Posición	**Frecuencia**
Point Guard	4787 (19 %)
Shooting Guard	5007 (20 %)
Small Forward	4881 (20 %)
Power Forward	5190 (21 %)
Center	4759 (19 %)
Total	**24624**

Cuadro 1: Distribución de las posiciones de los jugadores una vez realizadas las modificaciones.

[2]Se pueden consultar en el archivo, glosary.txt, adjunto en la subida del trabajo.

2.3.2. Asociaciones de variables

Correlaciones variables numéricas. Se ha calculado la correlación entre las variables numéricas, para ello se ha hecho uso de la función *cor* del paquete *stats* de *R*. Para el cálculo de las correlaciones se han filtrado los jugadores que han jugado más de 25 partidos en una temporada.

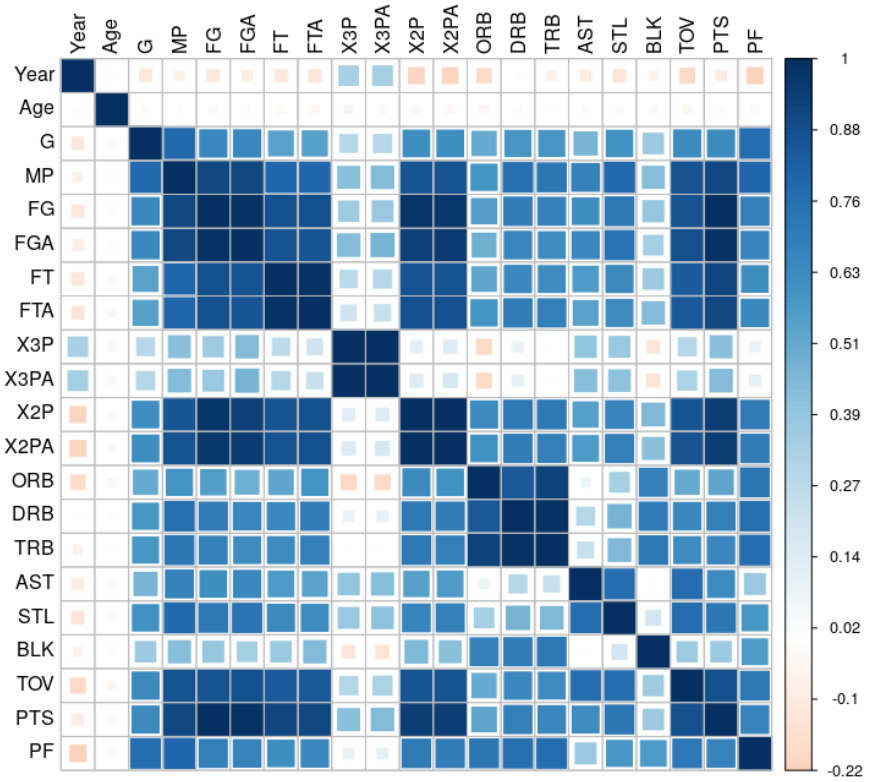

Figura 1: Correlación de las variables numéricas del conjunto original de datos. Las variables de menor correlación e interés han sido eliminadas para facilitar la lectura de la figura.

Se observan variables correlacionadas, algunos ejemplos esperables son la correlación de los puntos anotados con los lanzamientos o la relación entre rebotes ofensivos y defensivos. Sin embargo, por un lado resulta sorprendente que la variable edad del jugador no se correlacione con ninguna variable, era esperable que los jugadores más jóvenes y más veteranos tuviesen estadísticas más bajas. Así mismo sorprende que la variable año de la temporada tampoco se correlacione con relevancia, a excepción de los triples lanzados y anotados (a partir del año 1980 se incluyó la línea de triple), en la sección 2.3.3 se amplia el estudio sobre la evolución a lo largo de los años.

Correlación con variable categórica "Posición". Para observar la correlación de las variables con la asignada posición se ha hecho uso de la razón de correlación[3] y la función *eta2* disponible en los apuntes.

Posición	Variable	*eta2*
1º	TRB.	0.56
2º	AST.	0.53
3º	DRB.	0.47
4º	ORB.	0.36
5º	BLK.	0.29

Cuadro 2: Resultados de correlación con la clase usando la razón de correlación. Total de rebotes (TRB.), Asistencias (AST.), Rebotes defensivos (DRB.), Rebotes ofensivos (ORB.) y tapones (BLK.).

[3] http://fr.wikipedia.org/wiki/Rapport_de_corr%C3%A9lation

Las variables de rebotes (ofensivos, defensivos y totales) y tapones se asocian con la posición, parece lógico pensar que estas variables permiten identificar a los pivots. Por otro lado la variable de asistencias puede identificar a los bases. Sin embargo los valores obtenidos son bajos, sabiendo que la razón de correlación está en el rango [0,1] tan sólo las dos primeras variables superan 0,5.

2.3.3. Evolución durante los años

El conjunto de datos contiene información de más de 7 décadas distintas de juego, es previsible que durante los años el baloncesto haya evolucionado y por tanto las estadísticas sean diferentes entre décadas.

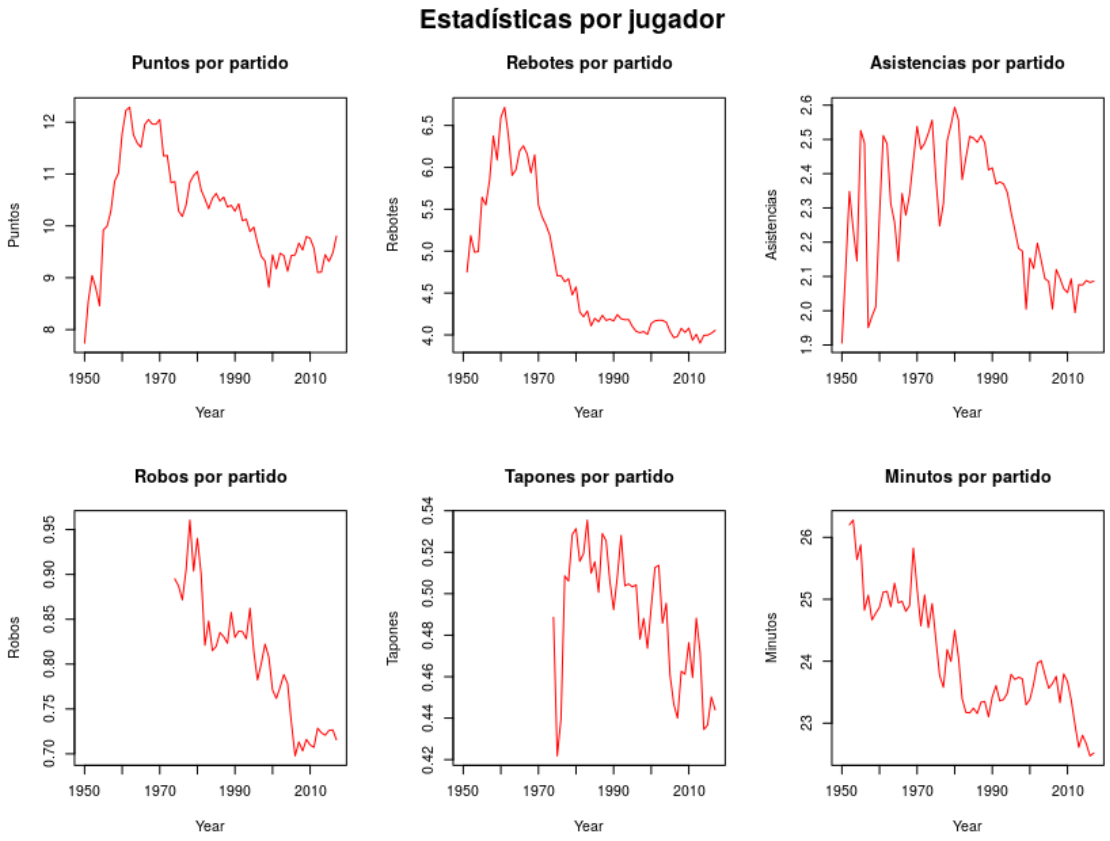

Figura 2: Evolución de puntos, rebotes, asistencias, robos, tapones y minutos por año. El eje y indica la media por jugador y por partido.

En la figura 2 se observa un descenso pronunciado en todas las estadísticas (excepto tapones) aproximadamente a partir de la década de los 70. La principal razón de este descenso se puede observar en la última gráfica, los entrenadores realizan más rotaciones de los jugadores por lo que juegan menos minutos, al jugar menos tiempo las otras estadísticas también se ven reducidas. Sin embargo si se calculan estas mismas gráficas con las medias por minuto el comportamiento es muy similar. Por lo tanto es lógico pensar que existe un cambio de tendencia en el juego, quizás se mejoran las defensas y existe mayor profesionalización de los jugadores. Finalmente es reseñable que a lo largo de los años la NBA se ha expandido habiendo más equipos y jugadores por equipo, por ejemplo en el año 1976 las dos ligas existentes (NBA y ABA) se fusionaron en la NBA.

Aunque las variaciones de las anteriores estadísticas son importantes, uno de los cambios más llamativo es la distribución de lanzamientos. Si bien hasta el año 1980 no se creó la línea de triple, en estas casi 4 décadas existe una evolución llamativa en la selección de tiro.

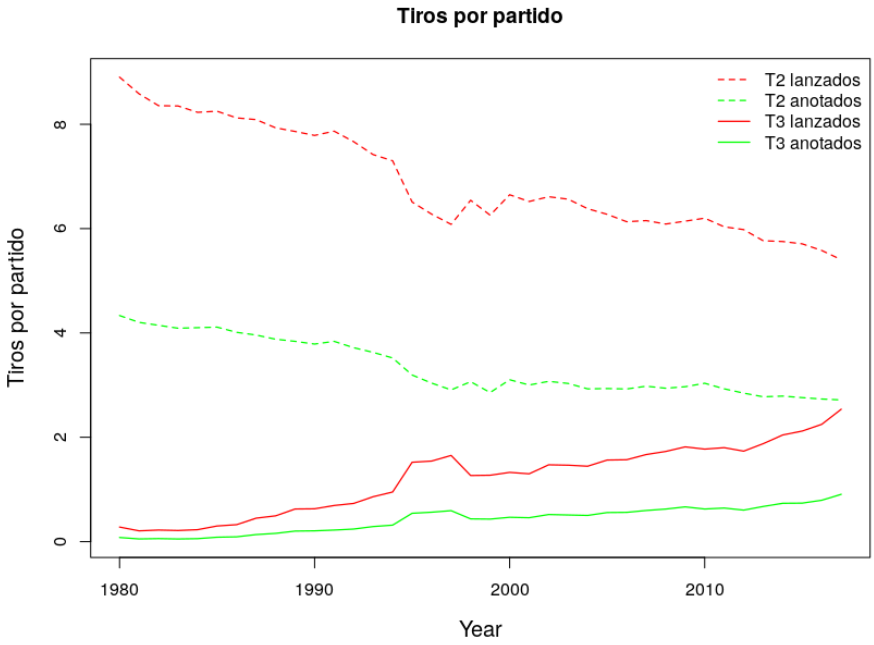

Figura 3: Evolución de la media por jugador de tiros por partido. Tiros de dos puntos (T2) y triples (T3).

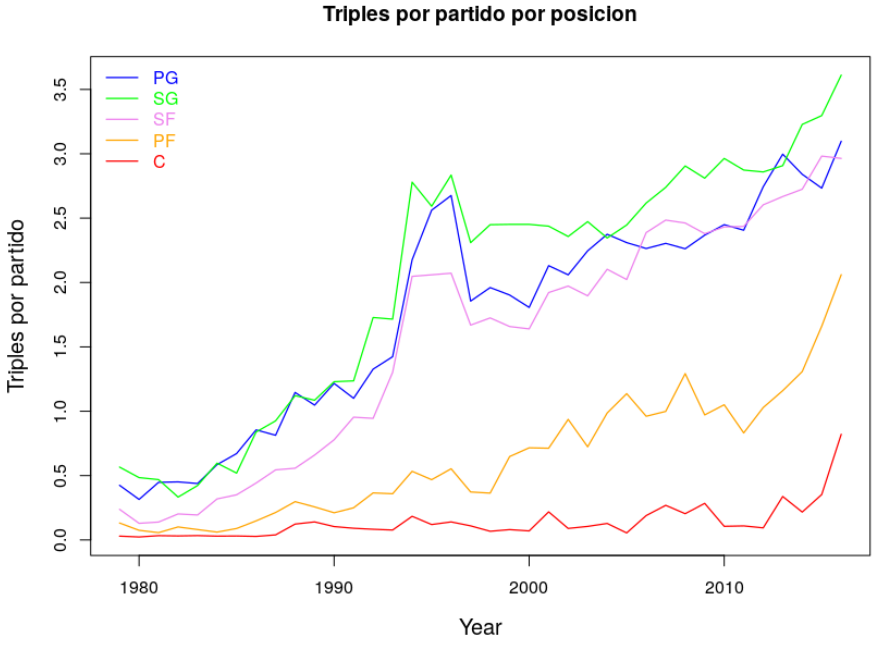

Figura 4: Evolución de la media por jugador de tiros por partido por posición. Base (PG), Escolta (SG), Alero (SF), Ala-Pivot (PF) y Pivot (C).

A lo largo de los años se ve una tendencia a tirar más triples, llegando a la actualidad a ser lanzados casi tantos triples como tiros de dos son anotados. Además la evolución según posiciones también es llamativa, aumentando considerablemente el número de triples lanzados por jugadores de mayor tamaño, Ala pivots (PF) y Pivots (C). Este aumento en los últimos años es muy pronunciado.

2.3.4. Estadísticas por altura

Una de las variables que puede ser más sencilla para determinar el estilo de juego de los jugadores es la altura, aunque existen jugadores que se salen de lo común es lógico pensar que el jugador más bajo será el base que se encargará de asistir a sus compañeros, en cambio el jugador más alto será pivot y será el encargado de rebotear y taponar.

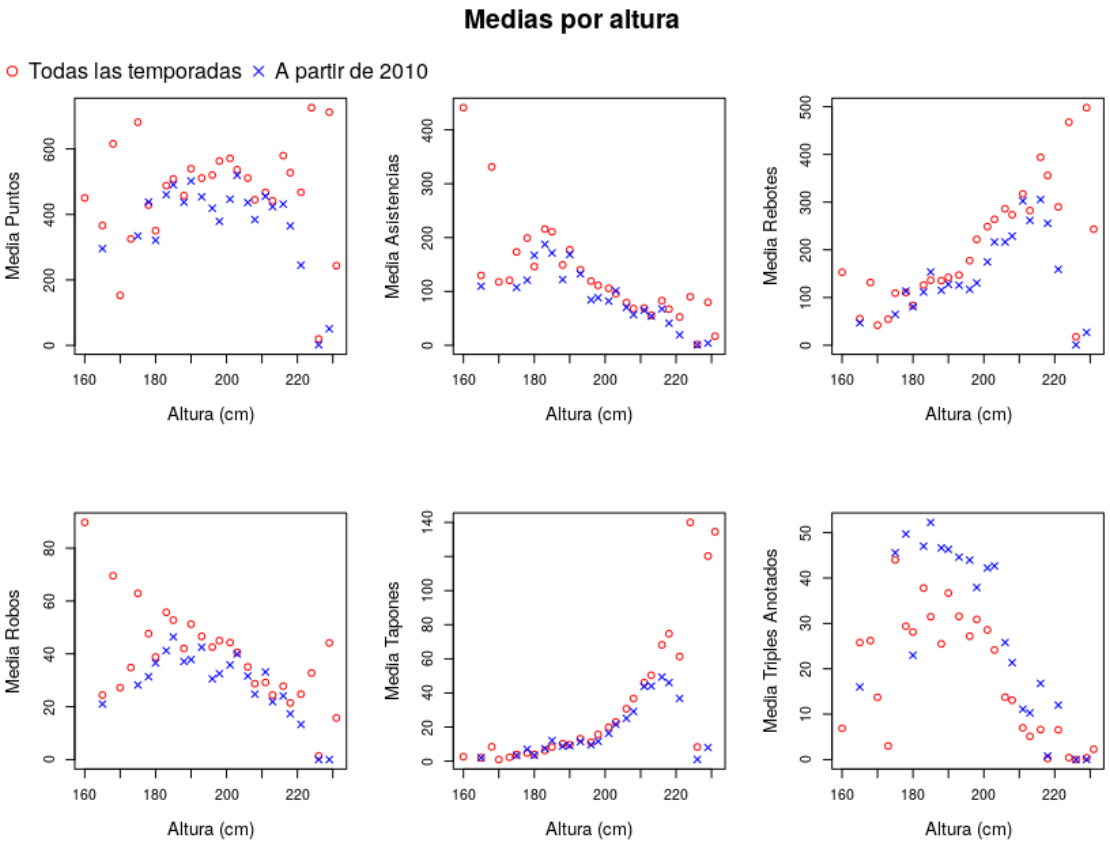

Figura 5: Medias de puntos, asistencias, rebotes, robos, tapones y triples anotados por temporada de los jugadores según altura.

Como se había previsto los jugadores más bajos asisten y roban más balones por temporada, en cambio los jugadores altos rebotean y taponan más. Esta tendencia ocurre tanto en las últimas temporadas como considerando todas las temporadas. La anotación está más distribuida, aunque se observa que en los últimos años los jugadores altos anotan menos.

2.3.5. Conclusiones

Parece claro que la evolución del baloncesto durante su historia ha sido bastante profunda, por tanto los jugadores de décadas pasadas poco tienen que ver con los actuales. Para las siguientes secciones se ha realizado un filtro por año para obtener jugadores contemporáneos que tengan unas estadísticas más homogéneas y faciliten la obtención de resultados.

Así mismo se ha detectado que la altura puede ser una variable de gran importancia para detectar jugadores de diferentes estilos de juego. En el dataset original de las estadísticas de las temporadas no se incluía la altura, se ha realizado un pequeño script para añadir por cada jugador su altura.

3. Principal Component Analysis (PCA)

3.1. Introducción

El análisis de componentes principales trata de reducir el espacio de variables realizando combinaciones lineales entre ellas [Abdi and Williams, 2010]. El conjunto del presente estudio presenta una gran oportunidad de aplicar dicha técnica al tratarse de variables que tienen gran dependencia unas de otras. Por un lado las variables *avanzadas* dependen directamente de las *tradicionales* y por otro lado como se ha observado existen relaciones entre las variables *tradicionales*. Por ejemplo los jugadores que cogen muchos rebotes también son los que más taponan, así mismo los jugadores que juegan más minutos también son los que más estadísticas totales tienen.

3.2. Aplicación

Se han realizado diferentes experimentos filtrando los datos originales en función de las variables, todas (52 variables) o las tradicionales (29 variables), y en función de las temporadas, todas las temporadas, a partir de 1979 (cuando se instauró el triple) y las del nuevo siglo (>1999).

Experimento	**Variables**	**Temporadas**	**Nº Comp (90 %)**	**Nº Comp (95 %)**
1	Todas (52)	Todas	15	21
2		>1979	14	19
3		>1999	14	19
4	Tradicionales (29)	Todas	9	13
5		>1979	9	12
6		>1999	8	11

Cuadro 3: Resultados PCA según el número de variables, temporadas y número de componentes para alcanzar 90 % o 95 % de la varianza.

Las dos últimas columnas de la tabla contienen el número de componentes necesarios para obtener el 90 % y 95 % de la varianza. Para todas las variables y en las últimas temporadas se podría reducir las 52 variables a 14 componentes manteniendo el 90 % o a 19 componentes manteniendo el 95 % de la varianza.

En el sexto experimento la primera componente supone más del 50 % de la varianza y unida a la segunda componente superan el 66 %.

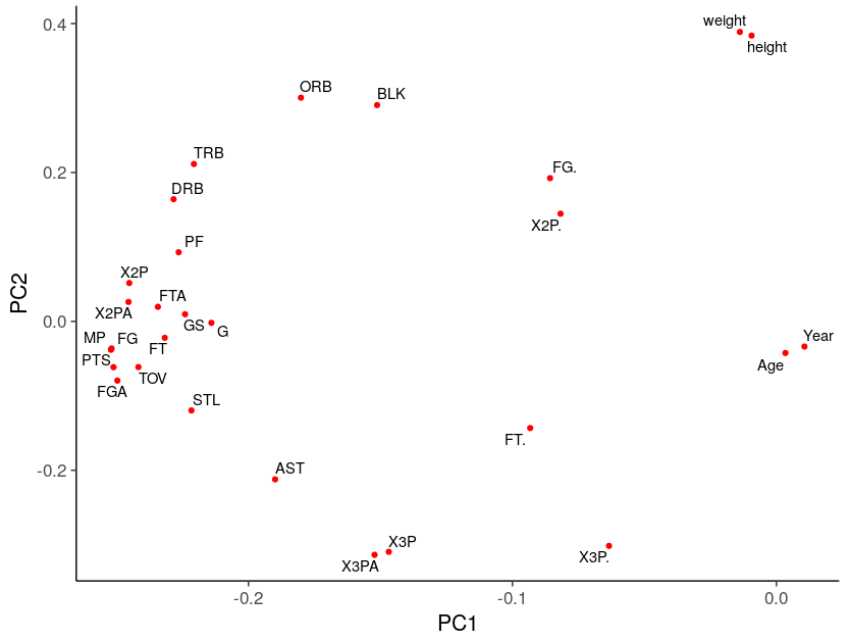

Figura 6: Representación de las dos primeras componentes del 6º experimento.

Se observa que en ambas componentes si existen valores altos se deberá a la altura y peso principalmente. En la primera componente los mayores valores también tendrán que ver con el año de la temporada y la edad del jugador, los valores negativos indicaran aspectos de anotación como los puntos anotados o tiros de dos puntos y tiro libres anotados. Por otro lado los valores negativos de la segunda componente se basan en las estadísticas de los triples.

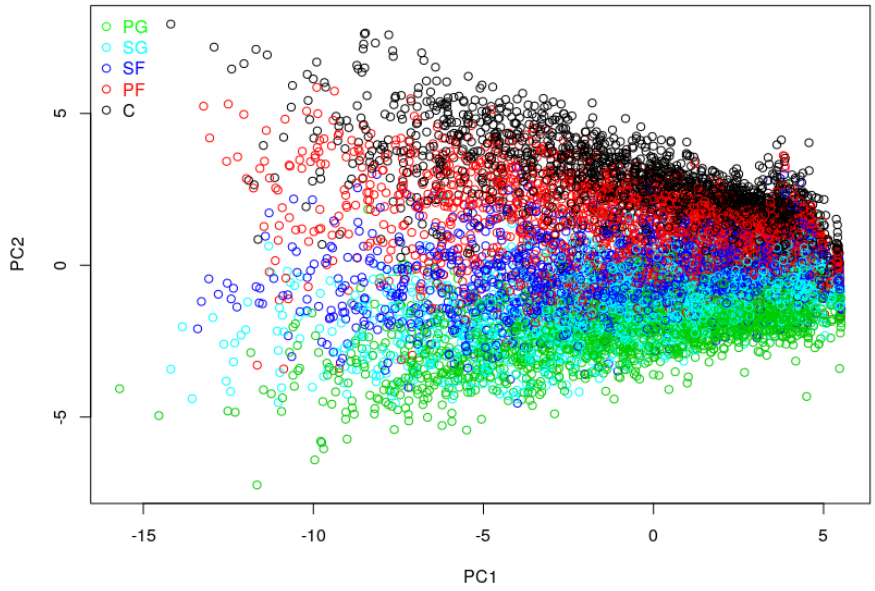

Figura 7: Representación de todos los jugadores según las dos primeras componentes del 6º experimento y coloreados según la posición original. Base (PG), Escolta (SG), Alero (SF), Ala-Pivot (PF) y Pivot (C).

Si representamos los jugadores según las dos componentes del sexto experimento y se colorean según la posición original se pueden observar diferentes capas según las posiciones. La posición más solapada es la de escolta (SG). Como se ha comentado en la figura 6 los valores negativos de PC2 se centraban en los triples, se observa que las posiciones predominantes situadas en la zona negativa de PC2 son bases (PG), escoltas (SG) y aleros (SF) siendo éstas las posiciones que más triples lanzan.

3.3. Conclusiones

Se observa que sobre este conjunto de datos la aplicación de *PCA* tiene un fuerte impacto, se podría reducir más de un tercio las variables tanto considerando todas las variables originales como sólo las *tradicionales* manteniendo una gran porcentaje de la varianza.

4. K-Means

Se ha realizado un estudio de clasificación sin supervisar, basándose en el algoritmo K-Means [Hartigan and Wong, 1979] para agrupar los jugadores de la última temporada (2017). Se pretende analizar si los cluster generados agrupan los jugadores según las posiciones originales del dataset o si proponen nuevas agrupaciones. Así mismo se ha realizado un estudio para detectar el número de K cluster optimo según el índice Silhouette [Laurentini, 1994] y el indice de la bondad disponible en los apuntes de la asignatura.

4.1. Búsqueda K óptimo

4.1.1. Silhouette

Se ha realizado un barrido sobre el parámetro K desde 2 hasta 15 para detectar la K que obtiene una mayor media del índice Silhouette. Se ha realizado considerando todas las variables numéricas originales del conjunto de datos, sólo las variables tradicionales y por último aplicando PCA y obteniendo los componentes que consiguen el 95 % de la varianza.

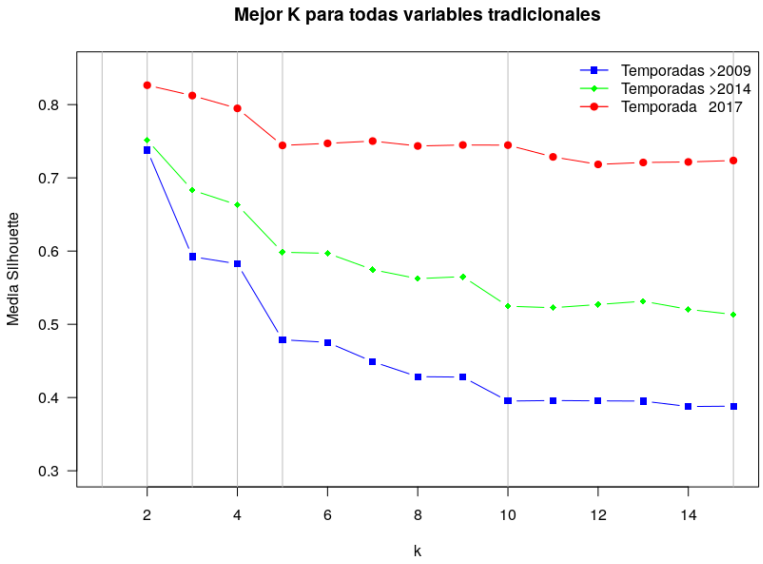

Figura 8: Mejor K considerando las variables tradicionales, para tres intervalos de tiempo diferentes.

Se observa que el K óptimo siempre es 2 y progresivamente según aumenta K la media del índice Silhouette disminuye, aunque sobre K=6-7 hay un ligero aumento. También se ha calculado la misma gráfica considerando todas las variables numéricas, los valores de la media de índice Silhouette son ligeramente inferiores y el comportamiento de decrecimiento de la gráfica según se aumenta K es casi idéntico a la figura 8.

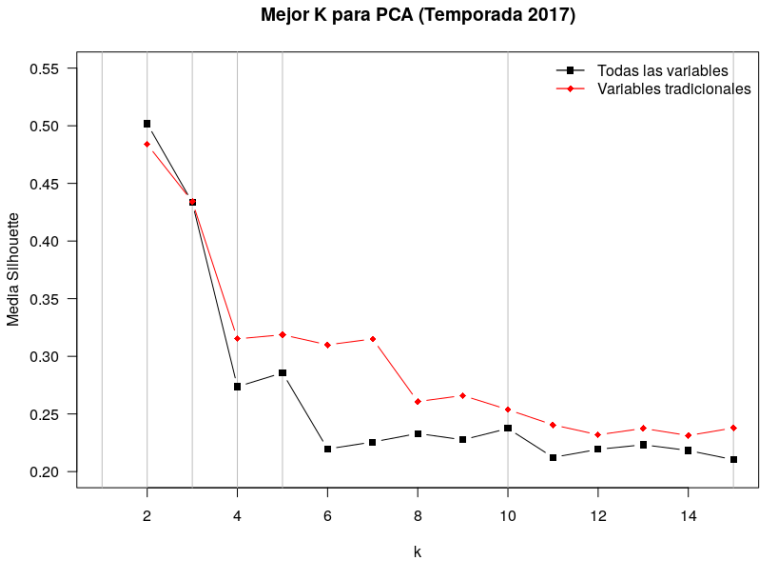

Figura 9: Mejor K haciendo uso de PCA considerando todas las variables y solo las tradicionales.

Finalmente si se aplica PCA, obteniendo 12 compontentes para 95 % de la varianza, el com-

portamiento es parecido siendo K=2 la mejor solución y decreciendo el valor del índice Silhouette según aumente K. Sin embargo, los valores de la media de Silhouette son más bajos que sin aplicar PCA. En la figura 8 los valores máximos se sitúan sobre 0,8 mientras que en la figura 9 están sobre 0,5.

4.1.2. Índice de Bondad

Se ha validado el mejor resultado de índice Silhouette, es decir con los datos de la temporada 2017 y variables tradicionales:

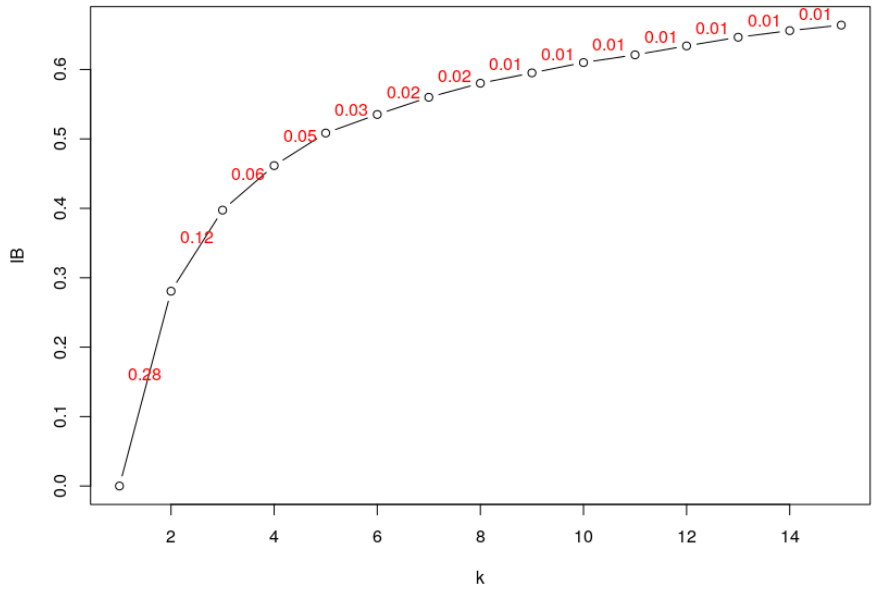

Figura 10: Validación de la mejor K haciendo uso del índice de bondad (IB).

Para valorar el índice de bondad hay que valorar los incrementos a medida que aumenta la K, se observan incrementos muy pequeños desde K=2, a partir de K=8 el incremento tan sólo es de 0,01.

4.2. Estudio clusters

Tras la búsqueda de la K óptima se ha obtenido que la mejor división posible entre clusters es K=2. A pesar de ser el que mejores valores de índice Silhouette y de bondad obtiene carece de interés ya que es conocido que existen más de dos tipos de jugadores. Por tanto en esta sección se analiza por qué K=2 es la mejor opción y se analiza como alternativa K=7, ya que sobre K=7 parece haber un máximo local con el índice Silhouette y en K=8 prácticamente el IB deja de crecer. Además 7 clusters es un número asimilable para estudiarlos individualmente.

4.2.1. Estudio 2 clusters

Se ha analizado la solución óptima con K=2, en la siguiente tabla se observa la distribución de posiciones originales:

Cluster	PG	SG	SF	PF	C	Total
1	36	54	38	31	25	184
2	19	11	18	15	20	83

Cuadro 4: Distribución de posiciones para K=2.

El primer hecho destacable es que el primer cluster contiene casi el doble de jugadores que el segundo. Por otro lado en ambos cluster la distribución de las posiciones de los jugadores es uniforme. Se ha calculado las medias de las variables entre los dos clusters y se obtiene que todas las medias, salvo la edad y porcentaje de triples, son superiores en el segundo cluster. Se puede considerar este segundo cluster como los jugadores *buenos*, es decir los que más puntos meten, más rebotes atrapan o más minutos juegan.

4.2.2. Estudio 7 clusters

Se ha analizado la solución alternativa con K=7, en la siguiente tabla se observa la distribución de posiciones originales:

Cluster	**PG**	**SG**	**SF**	**PF**	**C**	**Total**
1	1	1	8	19	13	42
2	16	27	11	3	1	58
3	0	0	0	0	7	7
4	23	31	23	15	2	94
5	1	0	6	8	20	35
6	12	6	6	1	0	25
7	2	0	2	0	2	6

Cuadro 5: Distribución de posiciones para K=7. Base (PG), Escolta (SG), Alero (SF), Ala-Pivot (PF) y Pivot (C).

La descripción de los clusters:

1. Especializado en posiciones interiores, ala-pivot y pivot, además de media tiran y meten más tiros de dos puntos, pero ligeramente tiran y anotan menos triples que los otros jugadores de esas mismas posiciones. También juegan más tiempo, cogen más rebotes y cometen más faltas personales. Algunos ejemplos: Blake Griffin, Marc Gasol o Pau Gasol.

2. Especializado en posiciones exteriores, base, escolta y alero. Son más bajos que la media de sus posiciones y engloban dos tipos de jugadores: triplistas (JJ Redick o Ryan Anderson) y defensores exteriores como (Avery Bradley o Marcus Smart).

3. Cluster puro donde sólo hay pivots. Además son los pivots que más rebotes, tapones y mejores porcentajes de tiro interior tienen. Por otro lado son los pivots menos triplistas. Ejemplos: Andre Drummond, Rudy Gobert, Dwight Howard o DeAndre Jordan.

4. Cluster bastante distribuido, salvo en los pivots, y además el que más jugadores contiene. Son los jugadores que menos anotan y juegan.

5. De nuevo cluster que agrupa en su mayoría pivots, en este caso son los que menos minutos juegan, anotan y rebotean. Ejemplos: Timofey Mozgov, Zaza Pachulia o Noah Vonleh.

6. Agrupa en su mayoría bases (PG) y jugadores exteriores, en concreto los que más anotan y más minutos juegan. Ejemplos: Stephen Curry, Kawhi Leonard o Kyrie Irving.

7. Tan sólo agrupa a seis jugadores de diferentes posiciones, estos jugadores son los que más anotan y grandes estrellas de la liga: James Harden, Anthony Davis, Lebron James, Giannis Antetokounmpo, Russell Westbrook y Karl-Anthony Towns.

4.3. Conclusiones

Como solución optima K=2 no presenta gran interés al ser demasiado pequeño, se ha propuesto una división entre 7 posiciones que difieren de las 5 posiciones originales. Se ha observado que la detección de jugadores altos (pivots) es la más sencilla, también la separación de los jugadores exteriores es más o menos buena, siendo los bases la posición más fácil de identificar.

Se han obtenido 5 clusters interesantes que permiten detectar jugadores de juego interior (cluster 1), exteriores especialistas en triples o defensa (cluster 2), mejores pivots defensivos (cluster 3), exteriores anotadores (cluster 6) y los mejores jugadores de la NBA (cluster 7).

Referencias

[Abdi and Williams, 2010] Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

[Hartigan and Wong, 1979] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[Laurentini, 1994] Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162.